Synopsis V6.0
Proton Single Event Effects (SEE) Testing of the Myrinet Crossbar Switch and Network Interface Card†

Jim Howard[1], Ken LaBel[2], Marty Carts[3], Ron Stattel[3], Charlie Rogers[3] and Tim Irwin[4]
1. Jackson and Tull Chartered Engineers, Washington DC 20018
2. NASA GSFC, Greenbelt, MD 20771
3. Raytheon ITSS, Greenbelt, MD 20771
4. QSS, Inc., Greenbelt, MD 20771

Test Dates: November 26-28, 2001. Report Date: January 31, 2002.

Introduction

As part of the Remote Exploration and Experimentation Project (REE), work was funded for the "Radiation Evaluation of the INTEL Pentium III and Merced Processors and Their Associated Bridge Chips." As a continuing step in the completion of this work, REE requested a proton SEE evaluation of the Myricom network protocol system (Myrinet) that included the Myrinet crossbar switch and the Network Interface Card (NIC). To this end, two crossbar switch devices and five components in the NIC were exposed to the proton beam at the University of California at Davis Crocker Nuclear Laboratory (CNL).

Test Facility

Facility: University of California at Davis Crocker Nuclear Laboratory
Protons: 63 MeV
Flux: 1.8×10^7 to 1.0×10^9 protons/cm^2/s.

Test Methods

Temperature:

The test was conducted at room temperature.

Devices Tested

Two devices tested were crossbar switches manufactured by Myricom, Inc, which provide the interconnectivity in the Myrinet model (as a hub would in star-configuration network model). In addition, a Peripheral Component Interconnect (PCI)-bus network interface card (NIC) manufactured by Myricom was tested. On the NIC, five devices were arbitrarily chosen to be exposed and the system evaluated for its response to their exposure. The listing of all devices used in this testing is given in Table I below. Official description of the Myrinet standard appears in its entirety in Appendix A. Pertinent aspects of Myrinet are described when appropriate within this text.

16 port Crossbar

The crossbar switch (Xbar) device type is the essential component interconnecting devices residing on the network. The 16-port Xbar switch tested allows 16 devices to connect in any configuration, one with another (no broadcast or group connections). The bandwidth of Myrinet is described as the data rate available in the "forward"

† This work was performed at NASA/GSFC for the Remote Exploration and Experimentation Project. This project is part of NASA's High Performance Computing and Communications Program, and is funded through the NASA Office of Space Sciences.

direction plus the bandwidth available in the opposite direction. There are different Myrinet speed standards. The Xbar and the NIC types tested are capable of operating at the Myrinet-2000 data rate of 2 gigabits per second (GBPS) in both directions simultaneously (full duplex). Thus, the data rate is expressed as 2000 + 2000. Each single-direction 2000 MBPS link is referred to as a channel. The opposite-direction pair of these channels is referred to as a link.

TABLE I
Device Under Test (DUT) Table

Device	Vendor	Location	Model Number	Serial Number	Other Part Markings
Switch 1	Myricom	Switch Board 1	M3-SW16-8S	84312	A-0041 0124
Switch 2	Myricom	Switch Board 2	M3-SW16-8S	87091	A-0041 0125
NIC	Myricom	NIC	M3S-PCI64B-2	90110	B-0111 0128
Lanai9	Myricom	NIC	9.1	0118	
SerDeSer	Myricom	NIC	1.1	0123	
PCIDMA	Myricom	NIC	1.3	0126	
Transceiver	Vitesse	NIC	VCS7146RH		0113LUBAD
SRAM	Samsung	NIC	K7N803601M		TKLB53BA KOREA TKG012DA KOREA

The Device Under Test (DUT) is a 0.25 μm commercial CMOS ASIC manufactured for Myricom to perform this function, but each DUT acquired was pre-mounted to a printed circuit board. Thus, the entire board was treated as the DUT. For incidental benefit other components on the DUT board were irradiated, separate from the Xbar Integrated Circuit (IC). Appendix B is the datasheet for the Xbar.

Each Xbar IC has 16 System Area Network (SAN) ports. The SAN specification for Myrinet is described in Appendix A. Briefly, SAN is a parallel data and control signals format for short haul (components no more distant than within one rack). Eight of these SAN links are brought to the front panel through a serializer/deserializer for connection to external components. The other eight ports are connected to the backplane connector for SAN connection to other components within the chassis that hold these cards. Providing 16 serial ports requires more than one Xbar card; the other eight backplane ports must be made available at the front panel. This is accomplished with a “spline” card, which does not contain an Xbar but merely converts the eight SAN channels to/from serial format. An inefficient arrangement of two Xbar cards can be used in place of one Xbar and one spline. A single Xbar card can be used if the eight backplane ports are not required. Some of this testing involved a single Xbar card and the rest involved two Xbar cards because the proton irradiation was penetrating enough to hit both Xbar ICs which are one above the other. This gave data on many more paths (described below in the DUT System description).

Messages are transported across a Myrinet as one or more packets. Packets are encoded with routing information that allows it to reach the desired destination. Each pass thru an Xbar (in a large network many Xbar transits may be required to reach the destination) involves one byte of routing information, which gives a relative (to the incoming port) output port. These routing bytes are removed as they are used, and

Cyclic Redundancy Check (CRC) is recalculated and appended so that the new packet (1 byte shorter) is correctly formatted. Packets that have inconsistent CRCs are simply dropped. This behavior is hard-wired within each component (within the LANai9 processor within the NIC, and within the Xbar IC). That is unfortunate for SEE testing—events are detectable only by their failure to arrive. No examination of erroneous data is possible.

No provisions for cooling were required for the Xbar.

Figure 1. The Myrinet Xbar card showing the backplane SAN ports at top and serial front panel ports.

Network Interface Card

The Network Interface Card (NIC) provides functionality for a device (PCI-bus computer, as are most desktops currently in use) to communicate via the Myrinet-2000 standard (See Appendix C). It is a PCI-64 form-factor card that can operate from 32 bit PCI busses also. It operates at either 5 V dc for PCI32 operation or 3.3V for either PCI32 or PCI64 operation. The card operates at either 33 MHz or 66 MHz PCI bus speed with ICs that provide bus interface (including PCI Direct Memory Access (PCIDMA)), protocol processing, and serialization and deserialization (SerDeser) functions. The primary focus was on the protocol processing IC, which is called the "LANai 9 processor".

Operating in typical PC situations (33 MHz PCI32 bus) allows a data rate of 132 MB/s. Operation at maximum data rate (66 MHz PCI64 bus) is still limited by either the host computer's capabilities or the PCI bus limits. 523 MB/s is the theoretical limit; the actual rate during testing was near that limit. To achieve full 2 GBPS rate 64 bit 200 MHz bus and high speed processors are required. Figure 3 shows the NIC. The additional signals for PCI-64 operation can be seen hanging to the right of the white PCI-32 socket. The black Myrinet cable can be seen at far left.

No provisions for cooling were required for the NIC.

Figure 2. Close-up of the Xbar IC. The total height above the board of the IC is 1.4 mm.

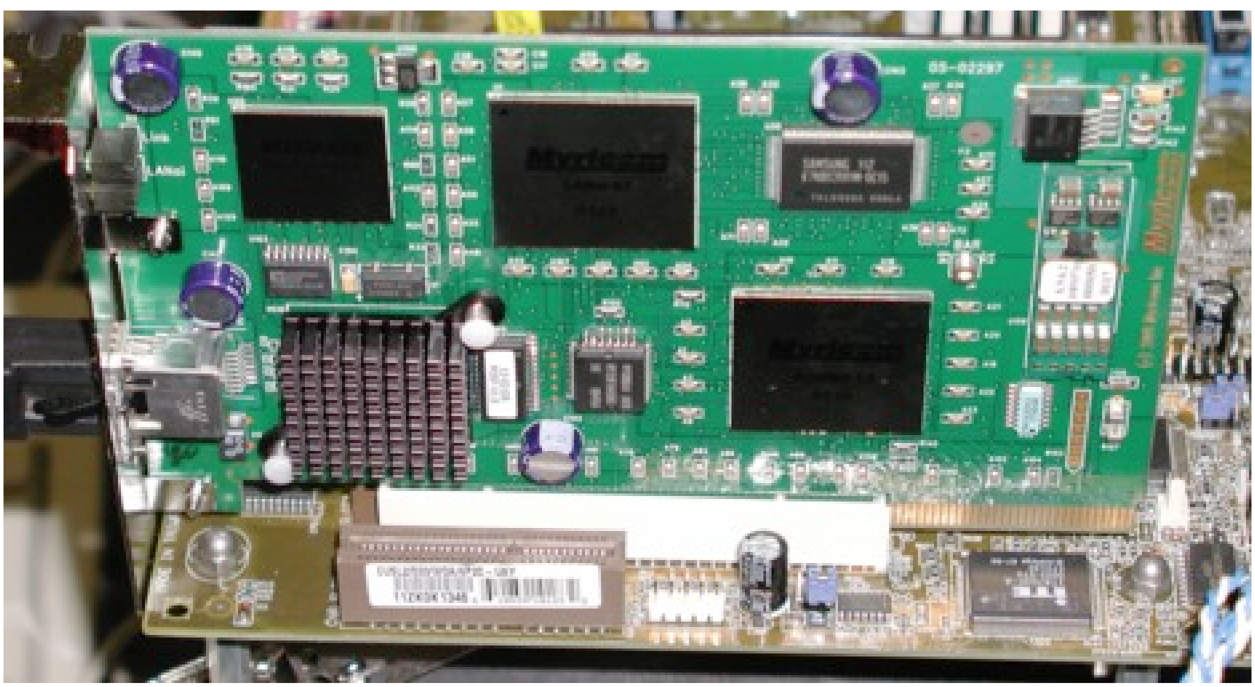


Figure 3. Myrinet NIC installed in PCI-32 motherboard without extender.

Test Hardware

The test system is substantially similar to the system used in related testing of Pentium III processors. The system hardware consists of two subsystems, the Test Controller and the DUT subsystems. Any cabling not required for the operation of the computer on which the NIC resides or the Xbar switch, along with Myrinet cabling is considered part of the Test Controller subsystem. These subsystems are described below. Figure 4 illustrates the overall test configuration.

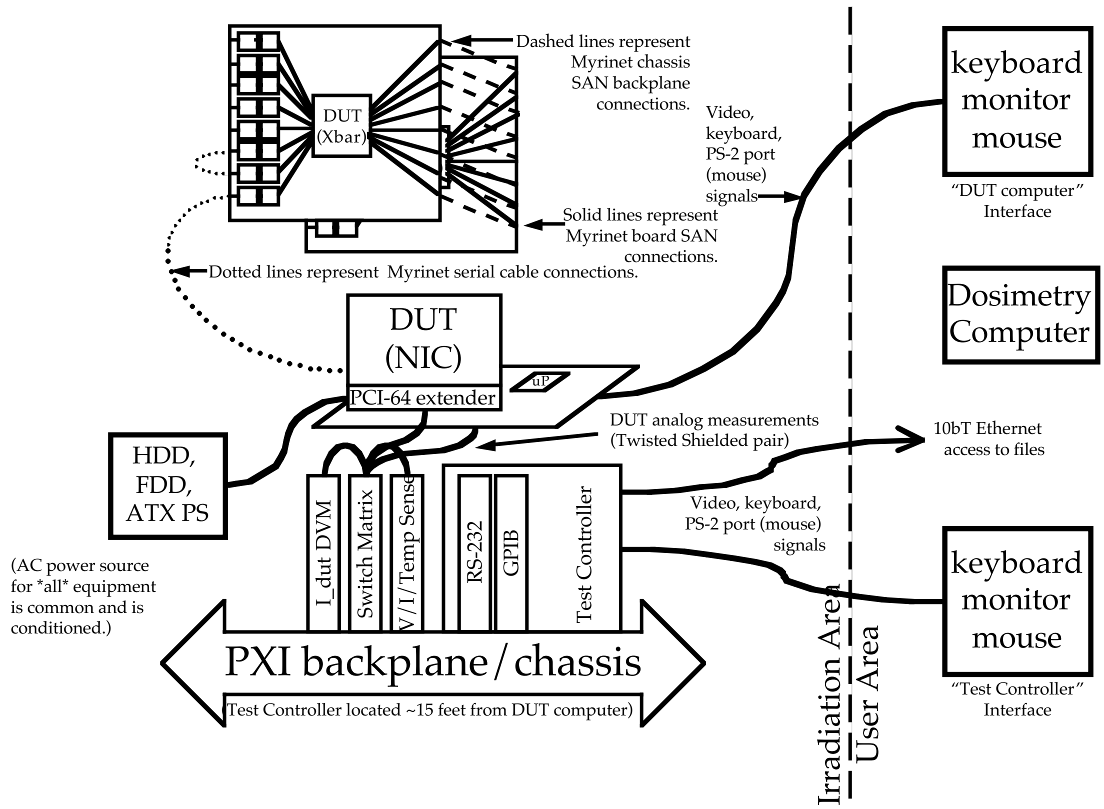


Figure 4. Block diagram of the test system.

Test Controller Subsystem

The Test Controller hardware is based on the PCI Extensions for Instrumentation (PXI) specification. The PXI subsystem, as shown in Figure 5, resides outside the irradiation area and is connected to the DUT at the irradiation point by cabling ~40′ long. It consists of the PXI components, the PXI Computer <-> DUT System cabling and the user interface.

The PXI components include the PXI chassis, which contains an embedded controller (running Win98, Labview™ (LV) environment and a custom LV application), a signal switch matrix, and two digital multimeters (DMMs) in the voltage measurement mode. The switch matrix provides two functions--The multiplexing of analog signals to one of the DMMs, and contact closures (pulling signal levels to ground). One DMM measures all analog values except the value read most frequently or as most important to not be delayed by switch settling time. The other DMM is dedicated to monitoring that value.

The PXI Computer's user interface, network connectivity (for data file access) and AC power feed are also components of the PXI Computer. An extended (via a CAT5 cable based extender from Cybex, Inc.) keyboard/monitor/mouse user interface provides user control of the PXI computer from the user facility (which in this instance is located in the hallway outside the regular, restricted access, user area).

Most of the PXI Computer <-> DUT System cabling leaves the PXI subsystem from the switch matrix (described further below). Exceptions are the AC power cable to power the DUT System and a serial (RS-232) cable for telemetry/command of the DUT System computer (telemetry originates within the DUT System computer, commands originate within the PXI Controller).

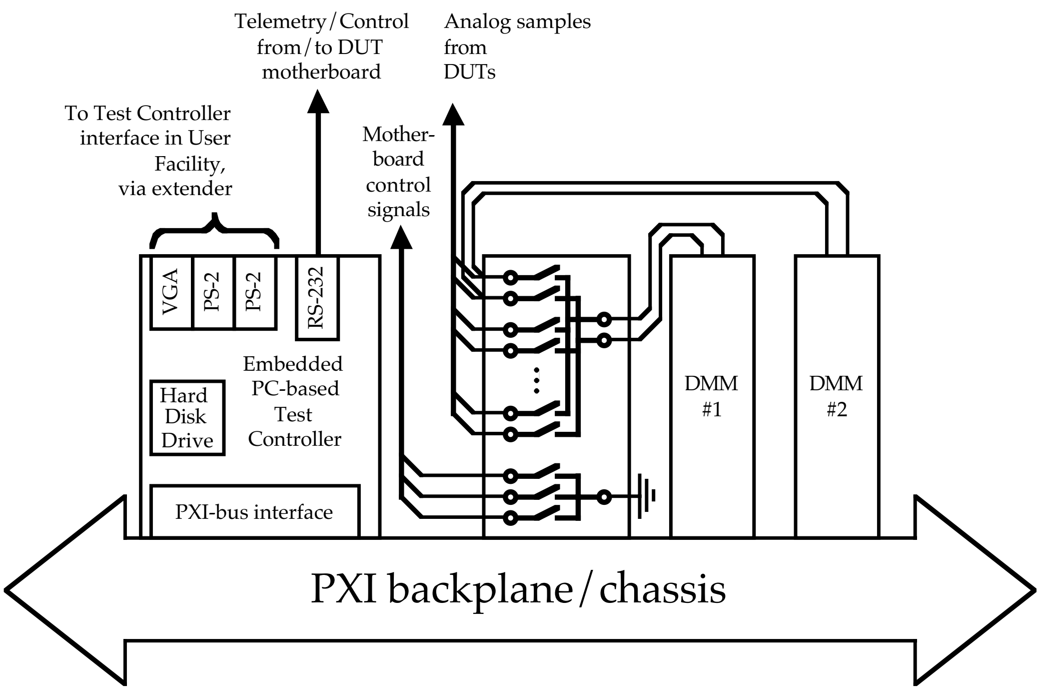


Figure 5. Block diagram of the PXI subsystem.

DUT System

The DUT System consists of the computer in which the NIC resides and to which the Xbar is connected. It includes components mounted directly to the motherboard, components located nearby (e.g. disk drives) and connected via cables, and a Cybex extended keyboard/monitor/mouse user interface.

The DUT system computer motherboard resides in the test chamber, positioned just below the particle beam when the NIC is exposed. The NIC plugs into an extension socket which raises it up by ~2". The dual 1 GHz P3 processors on the motherboard are Flip-Chip Pin Grid Array (FC-PGA) form-factor so they lay very low and well out of the particle beam, as do the low profile (<1") RAM modules.

Located nearby (~6 feet) are a modified standard PC ATX power supply (PS), a floppy and/or hard disk drive, and a Cybex user interface extension identical to the one used to extend the PXI computer.

The motherboard is modified to allow power cycling and reset via the PXI switch matrix. The ATX power supply is modified to allow force power shutoffs.

The PCI-64 extension board, which the NIC plugs into, is modified to sample DUT current via the PXI switch matrix and DMMs.

The NIC is connected via a Myrinet cable to the rest of the DUT system, the Xbar switches. (See Figure 6. Above and below angles of both identical ends are shown). These are housed in a chassis containing its own AC power supply.

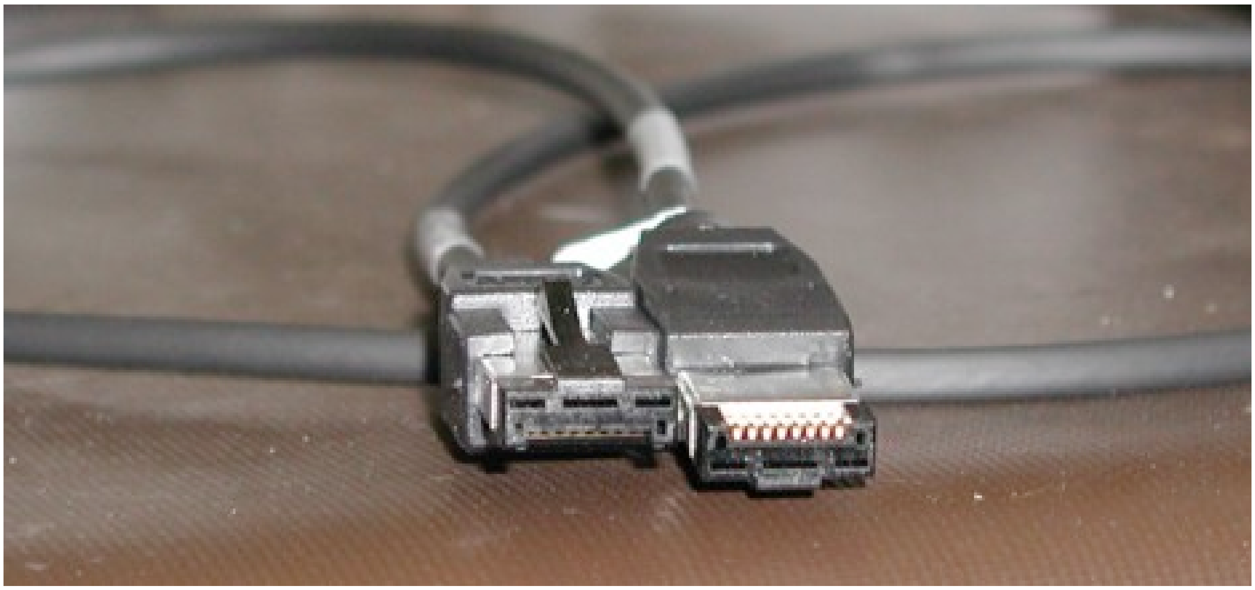


Figure 6. Close-up of Myrinet serial cable ends.

The motherboard is modified to allow connection to two controlling signals, both momentary contact closures. The motherboard front panel power on/off (MotherPonoff) input signal is controlled by the PXI switch matrix, as is the motherboard front panel soft reset (MotherSR) input signal.

ATX PS on/off state is normally controlled by a constant signal from the motherboard (The ATX SP supplies a standby +5V to power such motherboard functions). This signal (PS_ON#) is, approximately, a latched toggle of the front panel signal, MotherPonoff. This motherboard PS_ON# signal is disconnected from the ATX power supply's PS_ON# input so that that can be controlled directly from the PXI. This additional control is necessary because the computer can hang to the extent of not responding to the normal on/off commands. The ATX PS AC power is extended back to the user facility.

The DUT Computer runs the Windows-NT™ operating system and a software application that accesses Myrinet NIC drivers. Commands from the PXI computer are received via an ethernet cable and responses are transmitted back via the same link.

Currents and voltages from as many as three devices (one NIC and two Xbars) were monitored. System cabling was designed to conform to other testing in the REE program. That testing allowed four current/voltage samples in one subD 15-pin connector cable. A cable assembly was added to trifurcate three signals to separate locations.

DUT system signals that are controlled by the PXI subsystem, as described above, or by the user from the user facility are:

Name	Destination	Description
PS_ON#	ATX Power supply	Hold low (0 V) for PS on; Open = High = Off
MotherPonoff	Motherboard power switch connector	Pulse low (0 V) to toggle power on and off
MotherSR	Motherboard reset switch connector	Pulse low (0 V) to initiate reset
Command	DUT system computer	CAT-5 cable, ethernet, 10/100 mbps rate. Same cable that carries Telemetry data.
Keyboard/ mouse	DUT system computer	PS-2 keyboard ports

DUT computer signals that are monitored by the PXI or directly by the users in the user facility are:

Name	Source	Description
V_NIC, I_NIC	NIC extender card	Voltage and current samples of the NIC primary supply. Twisted shielded pair (TSP).
V_Xbar1, I_Xbar1	First Xbar card	Voltage and current samples of the first/only Xbar switch card supply. TSP.
V_Xbar2, I_Xbar2	Second Xbar card	Voltage and current samples of the second Xbar switch card supply, if installed. TSP.
Telemetry	DUT system computer	CAT-5 cable, ethernet, 10/100 mbps rate. Same cable that carries Command data.
GUI output	DUT system computer VGA card.	Video carrying output to the user facility.

Test Software

The DUT software for Myrinet testing was written in Microsoft C++ Professional version 6.0. It was designed to run in Windows 2000 Professional service pack 2. The driver for the Myrinet network adapter was GM 1.1. This driver was downloaded from the Myricom website (http://www.myri.com/).

The Network Interface Card (NIC) takes data packets from the driver and sends/receives the packets through the cables and network switches. The receive function of this card rejects data packets when errors are detected. The method used for detecting errors is a CRC check byte at the end of each packet.

The DUT Software sends packets with an incrementing packet # and data which is a function of the packet #. If the packet number/16 is odd then the data is a stream of bytes with the value hex 55; otherwise, it is a stream of bytes with the value hex AA. After each packet is sent, the program waits until either a packet is received or ~10 microseconds, whichever comes first. There are two physical setups supported. The first setup uses one NIC for both sending and receiving. The second setup uses two NICs, one for sending and one for receiving.

The DUT is connected through a TCP/IP socket to the test controller system where the test controller system acts as the host and the DUT acts as a client. The IP address and port used for the test controller connection are hard-coded. When not connected the DUT tries once every 3 seconds to make a connection. The DUT sends telemetry information to the test controller system and records the same telemetry to a file on the DUT hard disk drive. The telemetry consists of a stream of 4-byte long integers sent LSB first with the following format:

```
// The last byte of 4 is a data code. The table below shows the
//  definitions for each code:
// FF timestamp and beam info     xx xx yy FF
//    xx xx relative timestamp
//    yy 01 for beam on, 00 for beam off
// FE Error in data packet        xx xx yy FE
//    xx xx location within packet
//    yy data read
// FD Skipped Packet(s)           xx xx xx FD
//    xx xx xx Number of skipped packets
// FC Skipped Packet(s) (Large/-) 00 00 00 FC
//                                xx xx xx xx
//    xx xx xx xx Number of skipped packets
// FB Buffer overflow             00 00 00 FB
// FA Header                      AA AA AA FA
//                                aa aa aa aa
//                                tt tt tt tt
//                                rr rr rr rr
//                                rr rr rr rr
//                                rr rr rr rr
//                                rr rr rr rr
//                                ff ff ff ff
//                                ff ff ff ff
//                                ff ff ff ff
//                                ff ff ff ff
//                                ff ff ff ff
//                                ff ff ff ff
//                                ff ff ff ff
//                                ff ff ff ff
//                                pp pp pp pp
//    aa aa aa aa Ascii Version
//    tt tt tt tt Absolute Time Stamp
//    rr rr rr rr... Route information
//    ff ff ff ff... Filename
//    pp pp pp pp Packet Size
// F9 Reconnect                   00 00 00 F9
// <F8 Packet Number              xx xx xx xx
//    xx xx xx xx Packet Number
```

The DUT software utilizes two methods of data transfer. The first is the standard gm_send_with_callback(). The second methods of data transfer is the undocumented gm_raw_send_with_callback().

The standard method of transfer uses handshaking that ensures that the data is received without any detected errors before the send is completed. If any errors are detected the data is resent until the data is received without detected errors or a timeout of about a minute is reached. Before running the DUT Software in this mode, the GM utility program gm_mapper_service must be executed. This cannot be executed while the DUT software is running. In this mode the speed of data transfer can be set using gm_set_speed. This method was not used in testing at Crocker Nuclear Lab, Davis CA in November of 2001.

The undocumented method of transfer (Raw Mode) uses no handshaking. If the data is received with detected errors it is rejected by the NIC and is never seen by the user software. The user software detects when a packet is skipped or any errors that are not detected by the NIC are received. When packets are skipped, the packet number of the packet received after the skip and the number of skipped packets are recorded. When errors are found within a packet, the packet number, the locations within the packet and the actual values of the bytes in error are recorded.

The DUT software is controlled through buttons and checkboxes on the DUT console. All of these can be manipulated through the keyboard and mouse of the DUT computer. Some of these can be controlled through the TCP/IP connection by the test controller system. These are controlled from the test controller by sending a one-byte command to the DUT. The following can be controlled both by the DUT and by the test controller system (The values at the end are the values for the command byte from the test controller system):

1. The button called "Run" is pressed to start logging data and to start sending data when "Loopback" is checked. (1-Run)
2. The button called "End" is pressed to terminate logging and to stop sending data when "Loopback" is checked. (2-End)
3. The checkbox on the console called "Beam" is checked when the beam is on. (3-Check; 4-Uncheck)
4. The checkbox on the console called "Loopback" is checked when using one NIC and not checked when using two. (5-Check; 6-Uncheck)
5. The checkbox "Raw Mode" is checked when using the raw data transfer mode and not checked when using the standard transfer mode. (7-Check; 8-Uncheck)

The following can only be controlled from the DUT:

1. The "Route" button and edit box are used to set the route that the data takes through the switches when using the raw data transfer mode.
2. The "Packet Size" edit box is used to set the data packet size (4088 is the default, 4096 is the max).
3. The "Directory" edit box is used to select the directory into which the telemetry files are stored.
4. The "Suffix" edit box is used to select text which is appended to "Run*", where "*" is the run number, when forming the file name for the telemetry file. The run number is incremented each time the "End" button is pressed.

Test Methodology

In this simple test, the main objectives were to observe what effects would be induced by proton irradiation, with specific concern to latchup sensitivity of any parts. Therefore, to achieve these goals, the main devices of both the crossbar switch and the network interface card (NIC) were place in the proton beam. During their exposure, the DUT computer was running software that was generating data to be passed along the network and watching for the arrival of these packets of information. While no direct evidence of upsets was possible, as explained previously, if data within the packet was corrupted, the Myrinet hardware would drop the packet. The missing packet would then be noticed by the DUT software and recorded. This is the main type of error observed. During exposure of the NIC it was also possible to induce errors in the data stream once the NIC accepted a package as valid. The methodology and software were also in place to observe these types of errors as well.

The methodology flow was to place the device to be exposed in the beam, start the DUT and PXI software systems, turn the proton beam on, and, finally, observe the effects. The proton beam remained on until either a preset amount of fluence was achieved or a functional interrupt or latchup was observed. Initially, the preset fluence was set to a smaller amount due to the uncertainty in the total dose response of any of the devices. As the testing proceeded and the devices appeared to withstand the dose sufficiently well, preset fluences were set to levels that there was typically a functional interrupt prior to the preset fluence level being reached.

For the errors that were observed, the test software recorded all pertinent information about the errors, including the manner in which they were received (e.g., did a single packet get dropped or were a sequence of packets lost in a very short time span). For functional interrupts, as much information that could be gleaned from the test system was recorded. In some instances it was simply that the DUT computer rebooted while in others it was detailed information about which switch in the crossbar devices induced the interrupt. If any latchup current would have been observed, the device, the peak current seen at the device, and the functionality after the latch would have been recorded.

Results

Network Interface Card

Single Event Latchup

For the simple test being performed on this system, the NIC current was monitored for the entire board. Therefore, determination of a latchup event in an individual component would have to generate sufficient current to be observable above nominal NIC current. For all five components exposed to the proton beam on the NIC, no high NIC currents or destructive events were observed. There were events on all five devices that led to functional interrupts (to be discussed next). These events could possibly be produced via a high current condition in the respective part, as a power cycle of the DUT computer was required to reset after the interrupt. However, since no events were destructive, it is impossible to say that latchup did or did not play any role in these events.

Single Event Functional Interrupts (SEFI)

When any of the five devices were exposed to the proton beam, the DUT computer system would experience a SEFI event at some point. This could be seen as the DUT computer either freezing or initiating a self-reboot. In all instances observed for all five devices, a power cycle of the DUT computer that housed the NIC was required to regain functionality. The SEFI cross sections measured for the five devices are shown in the last column of Table II.

TABLE II

Part	Accumulated Dose (krad)	Upset Cross Section (cm^2)	SEFI Cross Section (cm^2)
Lanai9	59.2	6.81E-12	1.14E-11
SerDeSer	53.1	1.52E-11	5.07E-12
PCIDMA	45.7	2.94E-12	1.18E-11
Vitesse	50.7	4.03E-10	7.95E-12
Samsung	9.1	8.84E-11	7.37E-11

Single Event Upsets

As discussed in the software section, missed packets are the normal mechanism for errors to display themselves for the test setup used here. For all but the Samsung device, this is the upset mechanism that was observed. For the Samsung SRAM part, the second upset possibility arose. These are errors that are received that are not detected by the NIC. In other words, data correctly received by the NIC is processed by the Lanai9 processor and stored into the Samsung SRAM. While in this stored location, it is altered and that error is detected. The upset cross sections measured for the five devices are shown in the third column of Table II.

It should be noted that two of these Samsung SRAM parts are exposed during the testing (one on each side of the board). It is not clear from the Myricom documentation how much of both of these parts are used and if their usage is equal. Therefore, the cross section is left as a total cross section for both parts (not per bit or per device).

Total Dose

While total dose testing was not included in the requirements of this testing, proton dose is accumulated over the course of the test. No parametric measurements of the devices are feasible with this test setup. However, no functional loss or functional performance degradation was observed throughout this entire test. Therefore, it can be stated that the devices are total dose functionally survivable to at least the maximum proton dose during this test. These dose levels for the five devices on the NIC are shown in the second column of Table II.

Crossbar switch Device

Single Event Latchup

For the test being performed on this system, the crossbar switch current was monitored. For both crossbar switch devices exposed to the proton beam, no high currents or destructive events were observed. There were events that led to functional interrupts (to be discussed next). These events did require a power cycle of the crossbar power supply to reset after the interrupt. However, since no events were destructive and no high currents were observed for the switches, it is possible to say that latchup did not likely play any role in these events.

Single Event Functional Interrupts (SEFI)

Figure 7 and Figure 8 show the per-switch SEFI cross-section for the crossbar switch devices tested. In Figure 7, it is assumed that all switches have the same sensitivity whether they are on the in Xbar frontplane (FP) or backplane (BP), on either Xbar #1 or #2. This cross section is plotted as a function of the number of switches active during that test (there are different percentages of the switch locations for these four cases). The black squares and error bars are the average overall cross section, assuming all switches the same, and the one-sigma standard deviation. The blue triangles are the cross sections within each of the four cases. The four data points almost lie within one sigma.

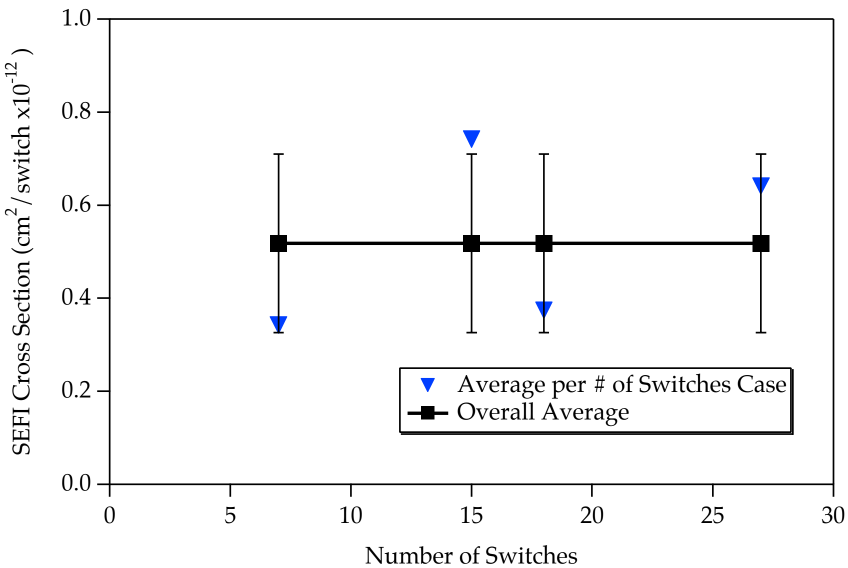


Figure 7. SEFI Cross section as a function of the number of switches.

Figure 8 looks at the same data set but with the thought that the SEFI rate could be different between frontplane and backplane switches. The four cases shown here are the two frontplanes of the two Xbars, the backplane switches, independent of which Xbar houses them, and the overall cross section (the same as the black squares of Figure 2). While all of the data points lie within the one-sigma error bars of the overall cross section, there does appear to be a difference between the frontplane and backplane switches.

Single Event Upsets – Non-SEFI

Single Event Upsets (SEU) for the crossbar switch devices are only evident as dropped packages. Data was collected about the number of dropped packages, including whether they arrived as a single dropped package or in a rapid sequence of dropped packages. This data was collected for four different switch quantities, that also had varying quantities in the frontplane and backplane.

Figure 9 shows the per-switch cross section as a function of the number of switches in the test configuration. It shows data for both single package loss and for multiple package loss. It is evident that the multiple package loss appears to be within approximately one sigma of an average value for the multiple events. The same can not be said for single package loss, as the two higher switch count cases (those with backplane switches in the test configuration) have substantially higher cross sections.

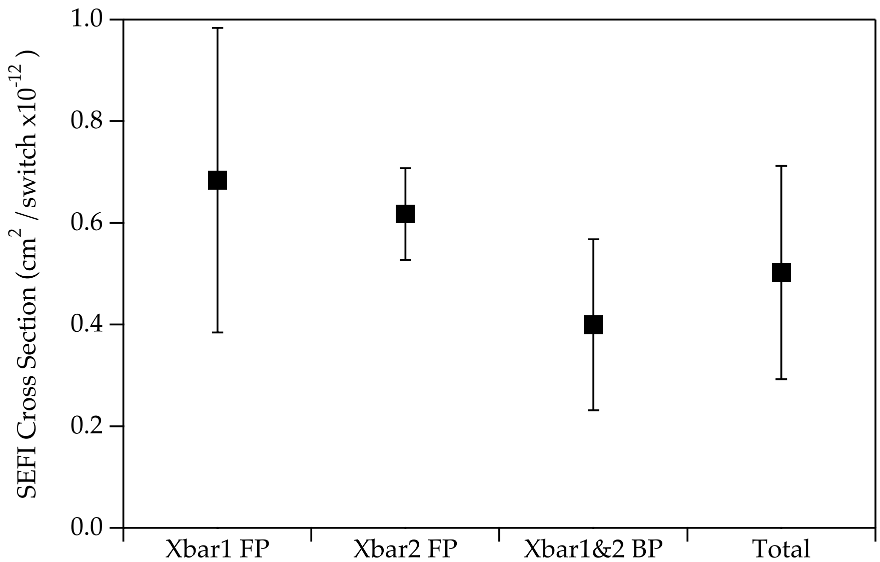


Figure 8. SEFI cross-section as a function of the location of the affected switch.

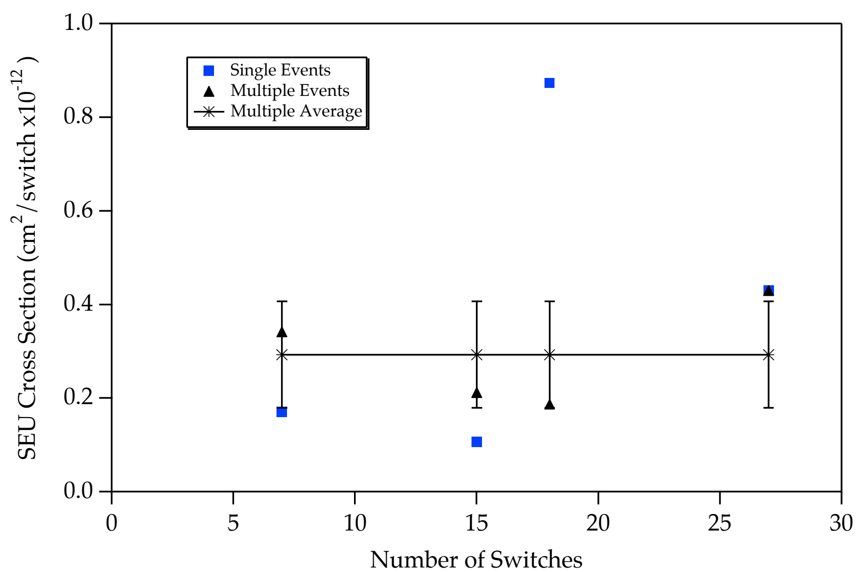


Figure 9. SEU cross section as a function of the number of switches for both single dropped packages as well as multiple dropped packages.

The same data as shown in Figure 9, can be viewed in another way by looking at the total cross section (both single and multiple package losses and not per-switch). This data is shown in Figure 10. For the two cases with the lowest number of total switches (the cases with only frontplane switches) have a cross section that is nearly an order of magnitude lower than the two cases with higher number of switches. Both of these higher switch count cases have the full sixteen switches from the backplane incorporated in the path for the data packages. The highest switch count case does have a slightly higher cross section than the next lower case as it does contain nine additional frontplane switches (again, these cross sections are not per-switch).

This SEU data appears to imply that having the backplane switches in the data path will substantially increase the data package loss as compared to running without backplane switches. It is possible that there are physical differences between backplane and frontplane switches in dealing with packages that is not immediately evident from the Myricom documentation.

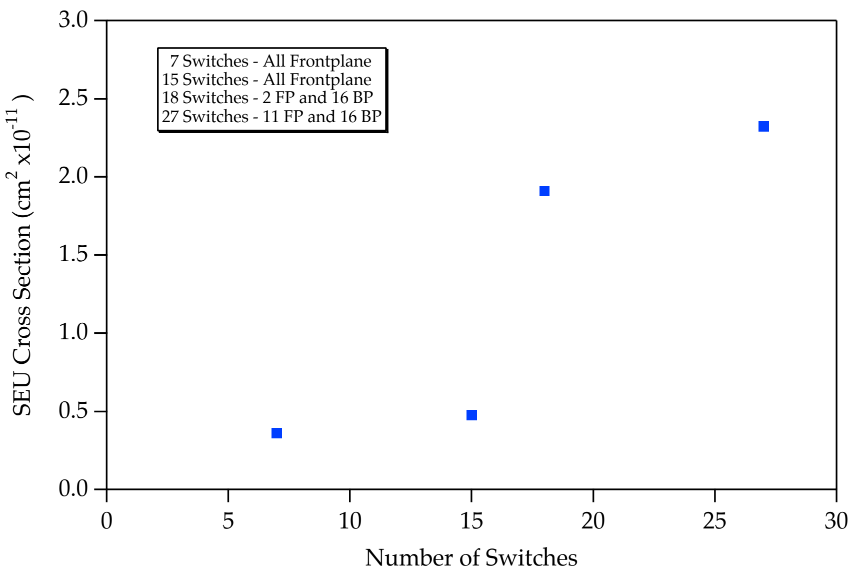


Figure 10. Total SEU cross section as a function of the number of switches with details of switch locations.

Total Dose

As with the NIC total dose testing was not included in the requirements of this testing. However, proton dose is again accumulated over the course of the test on the crossbar switch devices (Xbar). No parametric measurements of the devices are feasible with this test setup. However, no functional loss or functional performance degradation was observed throughout this entire test. Therefore, it can be stated that the devices are total dose functionally survivable to at least the maximum proton dose during this test. For this test setup, however, some amount of uncertainty exists for the dose levels of Xbar #2. This is because the proton beam passes through Xbar #1 and then the board for Xbar #1 before impinging on Xbar #2. While there is an unknown amount of material between the two switches, it does not appear to be substantial and it is assumed that the incremental doses on Xbar #1 are the same foe Xbar #2 when it is in place (Xbar #2 is only used when more than seven switches are used in the routing). These dose levels for the two crossbar switch devices tested are 400 krads and 285 krads, for Xbar #1 and #2 respectively.

Summary

The Myricom Myrinet network system was evaluated for proton single event effects response. No indication of latchup was observed. Functional interrupts and data loss upsets were observed and their cross sections determined.

ANSI/VITA 26-1998

Approved as an American National Standard by ANSI

American National Standard
for Myrinet-on-VME Protocol Specification

Secretariat
VMEbus International Trade Association

Approved November 2, 1998
American National Standards Institute, Inc.

VMEbus INTERNATIONAL TRADE ASSOCIATION
7825 E. Gelding Drive, Suite 104, Scottsdale, AZ 85260-3415
PH: 602-951-8866, FAX: 602-951-0720
E-mail: info@vita.com, URL: http://www.vita.com

ANSI/VITA 26-1998

American National Standard
for Myrinet-on-VME Protocol Specification

Secretariat
VMEbus International Trade Association

Approved November 2, 1998
American National Standards Institute, Inc.

Abstract
This standard describes a packet network protocol called Myrinet for communications between VME modules using interconnects either on a front panel or on a backplane. Networks may be module to module, subrack to subrack, and/or chassis to chassis.

American National Standard

Approval of an American National Standard requires verification by ANSI that the requirements for due process, consensus, and other criteria for approval have been met by the standards developer.

Consensus is established when, in the judgment of the ANSI Board of Standards Review, substantial agreement has been reached by directly and materially affected interests. Substantial agreement means much more than a simple majority, but not necessarily unanimity. Consensus requires that all views and objections be considered, and that a concerted effort be made toward their resolution.

The use of American National Standards is completely voluntary; their existence does not in any respect preclude anyone, whether he has approved the standards or not, from manufacturing, marketing, purchasing, or using products, processes, or procedures not conforming to the standards.

The American National Standards Institute does not develop standards and will in no circumstances give an interpretation of any American National Standard. Moreover, no person shall have the right or authority to issue an interpretation of an American National Standard in the name of the American National Standard Institute. Requests for interpretations should be addressed to the secretariat or sponsor whose name appears on the title page of this standard.

CAUTION NOTICE: This American National Standard may be revised or withdrawn at any time. The procedures of the American National Standards Institute require that action be taken periodically to reaffirm, revise, or withdraw this standard. Purchases of American National Standards may receive current information on all standard by calling or writing the American National Standards Institute.

Published by

VMEbus International Trade Association
7825 E. Gelding Dr., Suite 104, Scottsdale, AZ 85260

Printed in the United States of America - R1.0

ISBN 1-885731-15-9

1. General (Informative - Not Part of the Standard)

1. GENERAL (INFORMATIVE - NOT PART OF THE STANDARD)..................1

1.1 LIST OF TABLES...2
1.2 LIST OF FIGURES..2
1.3 LIST OF RULES..2
1.4 LIST OF RECOMMENDATIONS..4
1.5 LIST OF PERMISSIONS..4
1.6 LIST OF OBSERVATIONS...4
1.7 PURPOSE..5
1.8 SCOPE..5
1.9 TASK GROUP MEMBERS...6
1.10 VSO AND OTHER STANDARDS...7

2. INTRODUCTION TO THE MYRINET-ON-VME STANDARD.......................9

2.1 DEFINITIONS AND REFERENCES...9
2.2 REFERENCES...9
2.3 CONNECTOR NOTES..9
2.4 STANDARD TERMINOLOGY...9
2.5 THE STRUCTURE OF THE MYRINET-ON-VME STANDARD.........................10
2.6 GENERAL DESCRIPTION OF MYRINET *(NOT A PART OF THE STANDARD)*........12

3. SPECIFICATION OF THE DATA LINK LEVEL...............................15

3.1 MYRINET CHANNELS, LINKS, AND PORTS AND FLOW CONTROL..................15
3.2 MYRINET FLOW CONTROL...15
3.3 MYRINET PACKETS..16
3.4 MYRINET PACKET FORMAT..17
3.4.1 Myrinet Packet Header..17
3.4.2 Packet Payload...18
3.4.3 Packet Trailer...18
3.5 MYRINET COMPONENTS...19
3.6 MYRINET SOURCE ROUTES AND SWITCHING..................................20
3.7 MYRINET UNUSED / DISCONNECTED PORTS..................................21
3.8 MYRINET TOPOLOGY...21
3.9 TIMEOUT AND DEADLOCKS..22

4. THE REQUIREMENTS FOR THE PHYSICAL LEVEL............................23

4.1 MYRINET LINK RATES...23
4.1.1 640 Mbits/sec Rate...24
4.1.2 1280 Mbits/sec Rate..24
4.1.3 2560 Mbits/sec Rate..24

5. SPECIFICATION OF THE SAN PHYSICAL LEVEL............................25

5.1 THE SAN CHANNEL CHARACTER SET AND FLOW CONTROL.......................25
5.2 THE ENCODING OF THE SAN CHARACTERS ON SIGNALS........................27
5.3 THE CHARACTERISTICS OF SAN SIGNALS...................................29
5.3.1 Packet Timing..29
5.3.2 SAN Signal Characteristics...29
5.4 SAN CONNECTORS, CABLES, AND PIN-ASSIGNMENTS..........................30
5.4.1 Front-Panel and/or Rear-Panel Connection (SAN).....................30
5.4.2 Backplane Connection (P0/J0 and/or P4/J4)..........................33

6. SPECIFICATION OF THE LAN PHYSICAL LEVEL....35

6.1 THE LAN CHANNEL CHARACTER SET AND FLOW CONTROL....35
6.2 THE ENCODING OF THE LAN CHARACTERS ON SIGNALS....37
6.3 THE CHARACTERISTICS OF LAN SIGNALS....39
6.3.1 Packet Timing....39
6.3.2 LAN Signal Characteristics....39
6.4 LAN CONNECTORS, CABLES, AND PIN-ASSIGNMENTS....40
6.4.1 Front-Panel and/or Rear-Panel Connectors (LAN)....40

7. APPENDIX A: EXAMPLES OF MYRINET-ON-VME....43

7.1 EXAMPLES....43

8. APPENDIX B: EXISTING PRACTICE IN MYRINET ROUTING *(NOT A PART OF THE STANDARD)*....**45**

9. APPENDIX C: ASSIGNMENT OF MYRINET PACKET TYPES....47

10. APPENDIX D: GLOSSARY....49

1.1 List of Tables

TABLE 5–1: ENCODING OF SAN CHARACTERS....27
TABLE 5–2: SAN SIGNALING....29
TABLE 5–3: DUAL LINK SAN CONNECTOR PINOUT....31
TABLE 5–4: BACKPLANE CONNECTOR PINOUT....34
TABLE 6–1: ENCODING OF LAN CHARACTERS....37
TABLE 6–2: LAN SIGNALING....39
TABLE 6–3: LAN CONNECTOR PINOUT....40
TABLE 8–1: THE RELATIVE PORT ROUTING FOR A MYRINET 8-PORT SWITCH....46

1.2 List of Figures

FIGURE 2.5—1: THE MAP OF THIS DOCUMENT....12
FIGURE 3.1—1: MYRINET PORTS, CHANNELS, AND LINKS....15
FIGURE 3.4—1: THE STRUCTURE OF A MYRINET PACKET....17
FIGURE 7.1—1: FRONT PANEL INTERCONNECTION OF 4 SBCS....43
FIGURE 7.1—2: A BACKPLANE OVERLAY (WITH MYRINET/BP AND MYRINET/SAN)....43
FIGURE 7.1—3: FULL-BISECTION-BANDWIDTH 16-BOARD CONFIGURATION....44
FIGURE 7.1—4: 64 PROCESSORS IN A SUBRACK....44

1.3 List of Rules

RULE 3.1–1: MYRINET LINKS....15
RULE 3.1–2: CHARACTER SET....15
RULE 3.2–1: FLOW CONTROL....15
RULE 3.3–1: PACKET FRAMING....16
RULE 3.3–2: PACKET LENGTH (AND MTU)....16
RULE 3.4–1: PACKET STRUCTURE....17
RULE 3.4–2: HEADER STRUCTURE....17
RULE 3.4–3: VALID ROUTING BYTES....17
RULE 3.4–4: CRC-8....18
RULE 3.5–1: PROGRESS GUARANTEE....19
RULE 3.5–2: INVALID ROUTING....20
RULE 3.6–1: PACKET ROUTING BY SWITCHES....20
RULE 3.6–2: SOURCE ROUTING....20

RULE 3.7–1: DISCONNECTED PORTS 21
RULE 3.9–1: TIMEOUT 22
RULE 4.1–1: NOMINAL CHARACTER PERIOD FOR MYRINET-640....24
RULE 4.1–2: NOMINAL CHARACTER PERIOD FOR MYRINET-1280....24
RULE 4.1–3: NOMINAL CHARACTER PERIOD FOR MYRINET-2560....24
RULE 5.1–1: THE SAN CHARACTER SET 25
RULE 5.1–2: GAP SYMBOL....25
RULE 5.1–3: ASYNCHRONOUS CONTROL SYMBOLS....25
RULE 5.1–4: BEAT FREQUENCY 26
RULE 5.1–5: FLOW CONTROL 26
RULE 5.1–6: CABLE SLACK BUFFERS TO PREVENT DATA LOSS....26
RULE 5.2–1: SAN SIGNALS....27
RULE 5.2–2: ENCODING OF THE SAN CHARACTER SET 27
RULE 5.2–3: SAN NRZI ENCODING....28
RULE 5.2–4: SKEW 28
RULE 5.3–1: IPG (INTER-PACKET GAP)....29
RULE 5.3–2: SAN SIGNAL CHARACTERISTICS....29
RULE 5.3–3: COMMON GND FOR SAN SIGNALING 29
RULE 5.4–1: CONNECTOR GENDER....30
RULE 5.4–2: SAN CONNECTORS....30
RULE 5.4–3: THE A LINK IS THE PRIMARY LINK 30
RULE 5.4–4: UNUSED SAN LINKS....31
RULE 5.4–5: SAN PIN ASSIGNMENT....31
RULE 5.4–6: SAN CABLES 32
RULE 5.4–7: SAN CABLES FOR MYRINET-1280....32
RULE 5.4–8: SAN CABLES FOR MYRINET-2560....32
RULE 5.4–9: SAN CABLES – IMPEDANCE 32
RULE 5.4–10: CONFORMITY WITH VME64X 33
RULE 5.4–11: PIN ASSIGNMENT FOR THE BACKPLANE CONNECTORS 34
RULE 6.1–1: THE LAN CHARACTER SET 35
RULE 6.1–2: GAP SYMBOL....35
RULE 6.1–3: ASYNCHRONOUS CONTROL SYMBOLS....35
RULE 6.1–4: BEAT FREQUENCY 36
RULE 6.1–5: FLOW CONTROL (STOP AND GO SYMBOLS)....36
RULE 6.1–6: CABLE SLACK BUFFERS TO PREVENT DATA LOSS....37
RULE 6.2–1: LAN SIGNALS....37
RULE 6.2–2: ENCODING OF THE LAN CHARACTER SET....37
RULE 6.2–3: LAN DIFFERENTIAL SIGNALING 38
RULE 6.2–4: NRZI SIGNALING 38
RULE 6.2–5: SKEW 38
RULE 6.3–1: IPG (INTER-PACKET GAP)....39
RULE 6.3–2: LAN SIGNAL CHARACTERISTICS....39
RULE 6.3–3: COMMON GND FOR LAN SIGNALING 39
RULE 6.4–1: CONNECTOR GENDER....40
RULE 6.4–2: LAN CONNECTORS....40
RULE 6.4–3: LAN PIN ASSIGNMENT....40
RULE 6.4–4: LAN CABLE CONNECTIVITY 41
RULE 6.4–5: LAN CABLES 41
RULE 6.4–6: PIN #10 OF THE DB-37 CONNECTOR....41
RULE 6.4–7: LAN CABLES FOR MYRINET-640....41
RULE 6.4–8: LAN CABLES FOR MYRINET-1280....41
RULE 6.4–9: AUTO-SPEED FOR LAN PORTS....41

1.4 List of Recommendations

RECOMMENDATION 3.7–1: SELF-MONITORING....21
RECOMMENDATION 3.9–1: TIMEOUT PERIOD....22
RECOMMENDATION 5.1–1: BEAT SYMBOL (CONTINUITY MONITORING)....26
RECOMMENDATION 5.1–2: CABLE SLACK BUFFERS TO PREVENT DATA STARVATION....26
RECOMMENDATION 5.2–1: UNKNOWN CHARACTERS....27
RECOMMENDATION 5.2–2: SKEW....28
RECOMMENDATION 5.2–3: CABLE SKEW....28
RECOMMENDATION 5.3–1: PULLDOWN FOR INPUT SIGNALS....29
RECOMMENDATION 5.3–2: COMMON GND FOR SAN SIGNALING....29
RECOMMENDATION 5.4–1: POSITIONING OF SAN CONNECTORS....30
RECOMMENDATION 6.1–1: BEAT SYMBOL (CONTINUITY MONITORING)....36
RECOMMENDATION 6.1–2: CABLE SLACK BUFFERS TO PREVENT DATA STARVATION....37
RECOMMENDATION 6.2–1: UNKNOWN CHARACTERS....38
RECOMMENDATION 6.2–2: SKEW....38
RECOMMENDATION 6.2–3: CABLE SKEW....38
RECOMMENDATION 6.3–1: PULLDOWN FOR INPUT SIGNALS....39
RECOMMENDATION 6.4–1: TWISTED PAIRS....41

1.5 List of Permissions

PERMISSION 5.1–1: REPEATED GAP SYMBOLS....25
PERMISSION 5.2–1: SKEW REDUCTION....28
PERMISSION 5.4–1: RIGHT-SIDE-UP CONNECTORS....32
PERMISSION 5.4–2: RIGHT-SIDE-DOWN CONNECTORS....33
PERMISSION 5.4–3: SAN CONNECTOR LOCATION....33
PERMISSION 6.1–1: REPEATED GAP SYMBOLS....35
PERMISSION 6.1–2: REPEATED FLOW CONTROL SYMBOLS....36
PERMISSION 6.2–1: SKEW REDUCTION....38
PERMISSION 6.4–1: LAN CONNECTOR LOCATION....41

1.6 List of Observations

OBSERVATION 3.3–1: SETTING THE MTU....16
OBSERVATION 3.5–1: PROGRESS GUARANTEE....19
OBSERVATION 3.5–2: DROPPED PACKETS....20
OBSERVATION 3.6–1: SOURCE ROUTING....20
OBSERVATION 3.9–1: TIMEOUT PERIOD....22
OBSERVATION 4.1–1: MYRINET-RRR/PHY DESIGNATIONS....23
OBSERVATION 5.1–1: FLOW CONTROL APPLICABILITY....26
OBSERVATION 5.1–2: BYTES IN FLIGHT....26
OBSERVATION 5.2–1: SIGNAL FREQUENCY....28
OBSERVATION 5.4–1: DUAL LINK SAN CONNECTORS....30
OBSERVATION 5.4–2: SAN CABLE CONNECTIVITY....32
OBSERVATION 5.4–3: MYRINET-RRR/BP....33
OBSERVATION 5.4–4: CAUTION ON P0 AND P4 CONNECTORS....33
OBSERVATION 5.4–5: P0 HAS 7 ROWS (COPIED FROM VME64X OBSERVATION 4.6)....34
OBSERVATION 6.1–1: FLOW CONTROL ON OPPOSITE CHANNEL....36
OBSERVATION 6.1–2: FLOW CONTROL APPLICABILITY....36
OBSERVATION 6.1–3: BYTES IN FLIGHT....36
OBSERVATION 6.2–1: SIGNAL FREQUENCY....38

1.7 Purpose

The purpose of this specification is to provide the information needed to design VMEbus boards that interface with the Myrinet, high-performance, packet network.

Myricom Incorporated of Arcadia, California, developed the specification of Myrinet but makes no proprietary claims to this specification or to any technique. Myricom disclaims responsibility or liability for its use, or for any infringement of patents or other rights of third parties resulting from its use.

1.8 Scope

This standard describes the high-performance, inter-computer, Myrinet packet network that is fully compatible with existing VMEbus standards and their extensions. This standard addresses communication between VME boards using interconnect either on the front panel or on the backplane. The communication may use cables or an overlay (such as a backplane). The standard defines the interface between a VME board and Myrinet, allowing not only intra-subrack, board-to-board communication, but also a uniform extension for inter-subrack, inter-cabinet, and even local-area-network (LAN) communication.

This standard includes, either directly or by reference, the specification of the Data Link level, timing information, character set, signals, and the details of the connectors.

1.9 Task Group Members

When this standard was drafted, the Task Group for this standard had the following membership:

	Name		Company	E-address
1	Gorky Chin		Vista Controls Corporation	gorky@vistacc.com
2	Elwood Parsons	B	AMP	etparson@amp.com
3	J. J. Dumont		Framatome Connectors France	jdumont@iway.fr
4	David Wright		Hybricon Corporation	davidw@hybricon.com
5	Holly Sherfinski	B	Harting Inc. of North America	Holly.Sherfinski@harting.com
6	David Robak	B	Harting Inc. of North America	David.Robak@harting.com
7	Richard Jaenicke	B	Sky Computers, Inc.	jaenicke@sky.com
8	Istvan Vadasz		Force Computers	isva@force.de
9	Michael Thompson	B	Schroff, Inc.	m_thompson@ids.net
10	Martin Blake	B	VERO Electronics	martin_blake@vero-uk.com
11	Richard Morgan		Spectrum Signal Processing	rick_morgan@spectrumsignal.com
12	John Bratton		VERO Electronics	john_bratton@vero-usa.com
13	Danny Cohen	B	Myricom, Inc.	cohen@myri.com
14	Jim Waggett	B	CSPI, Inc.	jwaggett@cspi.com
15	Robert McKee		MITRE Corporation	rmckee@mitre.org
16	Michael Munroe		ERNI Components Inc.	mmunroe@compuserve.com
17	Wade Peterson	O	Consultant to Industry	peter299@maroon.tc.umn.edu
18	Douglas Endo	B	Raytheon System Company	dendo@msmail4.hac.com
19	Jing Kwok		Nexus Technology, Inc..	
20	John Rynearson	B	VITA	techdir@vita.com
21	Ray Alderman	B	VITA	exec@vita.com
22	Harry Andreas	B	Raytheon System Company	handreas@msmail4.hac.com
23	Joe Bedard	B	Hewlett-Packard Co.	bedard_j@apollo.hp.com
24	Tony Lavely	B	Mercury Computers, Inc.	atl@mc.com
25	Bob Patterson	B	AMP	rapatter@amp.com

The “B” in the third column indicates participation in the ballot group.
The “O” in the third column indicates Observers.

The following companies were sponsors of the development of this standard within the VSO (VITA Standards Organization).

AMP
CSPI
Myricom

When the ANSI Standards Board approved this standard on November 2, 1998, the balloting committee had the following membership:

Ray Alderman	VITA	<exec@vita.com>
Harry Andreas	Raytheon Systems Company	<haandreas@mail.hac.com>
Danny Cohen	Myricom, Inc.	<cohen@myri.com>
Adrian Cox	Transtech	<apc@transtech.co.uk>
Doug Doerfler	Sandia National Labs	<dwdoerf@sandia.gov>
Richard Jaenicke	Sky Computers, Inc.	<jaenicke@sky.com>
James Judd	Lockheed Martin Corp.	<jjudd@motown.lmco.com>
Elwood Parsons	AMP	<etparson@amp.com>
John Rynearson	VITA	<techdir@vita.com>
Michael Stern	CSPI	<mstern@cspi.com>
Rich Tallarico	Lockheed Sanders	<rtallari@sanders.com>
Michael Thompson	Schroff, Inc.	<m_thompson@ids.net>
James Waggett	CSPI	<jwaggett@cspi.com>

1.10 VSO and Other Standards

For information on other standards being developed by VSO, VME Product Directories, VME Handbooks, or general information on the VME market, please contact the VITA office at the address or phone number given on the front cover.

[This page intentionally left blank.]

2. Introduction to the Myrinet-on-VME Standard

2.1 Definitions and References

See Appendix D for the glossary of the new terminology specific to the added features described in this standard. Some of terminology described in this standard is repeated there.

2.2 References

The following publications are used in conjunction with this standard.

IEC 61076-4-101	IEC Standard for Hard Metric 2 mm Connectors
ANSI/VITA 1-1994	VME64 Standard, Approved April 10, 1995
ANSI/VITA 1.1-1997	VME64 Extensions (VME64x) Standard, Approved October 7, 1998.
ANSI/VITA 1.3-1997	ANSI/VITA 1.3-1997, VME64x 9U x 400 mm Format
ISO 7498	ISO Reference Model (ISORM) for Open Systems Interconnection, October 1984
NRZI	Chapter 19 by David R. Brown and Jack I. Raffel in Handbook of Automation, Computation, and Control, Ed. Grabbe, Ramo, and Wooldridge, John Wiley & Sons, 1959.

2.3 Connector Notes

The front-panel application uses SAN (System-Area Network) connectors that are in the Microstrip family. This family of controlled-impedance connectors can be used as board-to-board connectors as well as board-to-cable connectors.

Backplane applications use either P0 or P4 (as defined by ANSI/VITA 1.3-1997) as specified by IEC 61076-4-101. It is a 95 pin connector (P0/J0) that fits between the VME64x P1/J1 and P2/J2 connector pairs or the same connector (P4/J4) that fits between P2/J2 and P3/J3 or between P2/J2 and P5/J5. The 95 pin connector is a 2 mm Hard Metric style connector.

The interconnection of the VME environment to external environments is based on LAN cables that use low-voltage differential signaling (LVDS) over twisted-pairs, and use DB-37 connectors.

2.4 Standard Terminology

To avoid confusion and to make very clear what the requirements for compliance are, many of the paragraphs in this standard are labeled with keywords that indicate the type of information they contain. The keywords are listed below:

Rule
Recommendation
Suggestion
Permission
Observation

Any text not labeled with one of these keywords describes the Myrinet structure or operation. It is written in either a descriptive or narrative style. These keywords are used as follows:

Rule:
Rules form the basic framework of this draft standard. They are sometimes expressed in text form and sometimes in the form of figures, tables or drawings. All rules shall be followed to ensure compatibility between designs. All rules use the "shall" or "shall not" words to emphasize the importance of the rule. The "shall" and "shall not" words are reserved exclusively for stating rules in this draft standard and are not used for any other purpose.

Recommendation:
Whenever a recommendation appears, designers would be wise to take the advice given. Doing otherwise might result in some awkward problems or poor performance. While the Myrinet-on-VME architecture has been designed to support high-performance systems, it is possible to design a system that complies with all the rules but has abysmal performance. In many cases a designer needs a certain level of experience in order to design boards that deliver top performance. Recommendations found in this draft standard are based on this kind of experience and are provided to speed the learning curve. The "should" and "should not" words are reserved exclusively for stating rules in this draft standard and are not used for any other purpose.

Suggestion:
A suggestion contains helpful but not vital advice. The reader is encouraged to consider the advice before discarding it. Some design decisions are difficult without prior experience. Suggestions are included to help a designer who has not yet gained this experience. Some suggestions pertain to designing boards that can be easily reconfigured for compatibility with other boards, or to designing boards to ease system debugging.

Permission:
In some cases a rule does not specifically prohibit a certain design approach, but the reader might be left wondering whether that approach might violate the spirit of the rule or whether it might lead to some subtle problem. Permissions reassure the reader that a certain approach is acceptable and will cause no problems. The lower case word "may" is reserved exclusively for stating permissions in this draft standard and is not used for any other purpose.

Observation:
Observations do not offer any specific advice. They usually follow naturally from what has just been discussed. They spell out the implications of certain rules and bring attention to things that might otherwise be overlooked. They also give the rationale behind certain rules so that the reader understands why the rule must be followed.

2.5 The Structure of the Myrinet-on-VME Standard

This document specifies the Myrinet-on-VME Standard. It is written principally as an exposition, but interspersed with rules written in the restrictive or formal language of specifications. In addition, terms with formal meanings are shown in italics when first introduced and defined.

Section 2.6 describes Myrinet, in general, and its operation.

Chapter 3 specifies Myrinet at Level 2 of the ISO Reference Model (ISORM) for Open Systems Interconnection. This is the Data Link layer, the low level logical protocol. This chapter specifies also the abstract requirements that the Data Link layer imposes on the Physical layer underneath.

The Data Link level is the most appropriate starting point for the Myrinet specifications. Myrinet may be supported by various Physical-level (level 1 of the ISORM) implementations, for example, with electrical cables or a backplane, each with their own specified characteristics. At the Data Link level, however, Myrinet links are specified in more abstract terms that apply to all of the Physical-level implementations. Similarly, networks such as ethernet, ATM, and FDDI are defined at the Data Link level, and have multiple Physical-level implementations. The Data Link level also provides the foundation for higher protocol levels. The constancy of the Data Link specification, even when the Physical-level technology may change, protects investments in software at these higher levels. Chapter 3 defines the abstract Myrinet.

Chapter 4 summarizes the requirements that any Physical level implementation of Myrinet must meet, and defines the data rates of 640, 1,280 and 2,560 Mbits/sec.

Chapter 5 and Chapter 6 define the SAN (System-Area Network) and the LAN (Local-Area network) Physical levels of Myrinet-on-VME respectively. SAN implementations are intended for intra-cluster applications (including intra-PCB, intra-subrack, intra-cabinet, and even inter-cabinet) whereas the LAN implementations are intended for inter-cluster applications.

Both Chapter 5 and Chapter 6 specify for their types of networks:

(1) The channel character set (data bytes and control symbols)
(2) The encoding of the characters on signals
(3) The electrical characteristics of these signals (including rate, timing, and voltage)
(4) Pin-assignments, connectors, and cables.

The inter-board SAN links are expected to use either the front panel or the backplane. The LAN links are expected to run either from front-panel or from rear-panel connectors for high-performance communication with other clusters.

Chapter 5 specifies the standards for Myrinet-1280/SAN, Myrinet-2560/SAN, Myrinet-1280/BP, and Myrinet-2560/BP. Chapter 6 specifies the standards for Myrinet-640/LAN, and for Myrinet-1280/LAN

The above designations of the Myrinet standards include both the speed of the implementation and some reference to the implementation method (similar to the definition of the various ethernet variants such as E10, E100, E1000, E100BaseT, and E1000BaseT).

Appendix A shows several examples of Myrinet-on-VME. Appendix B describes the existing practices in Myrinet routing. Appendix C lists the assignment of Myrinet packet types, as of January 1998. Appendix D provides a glossary.

The following is a “map” of this document:

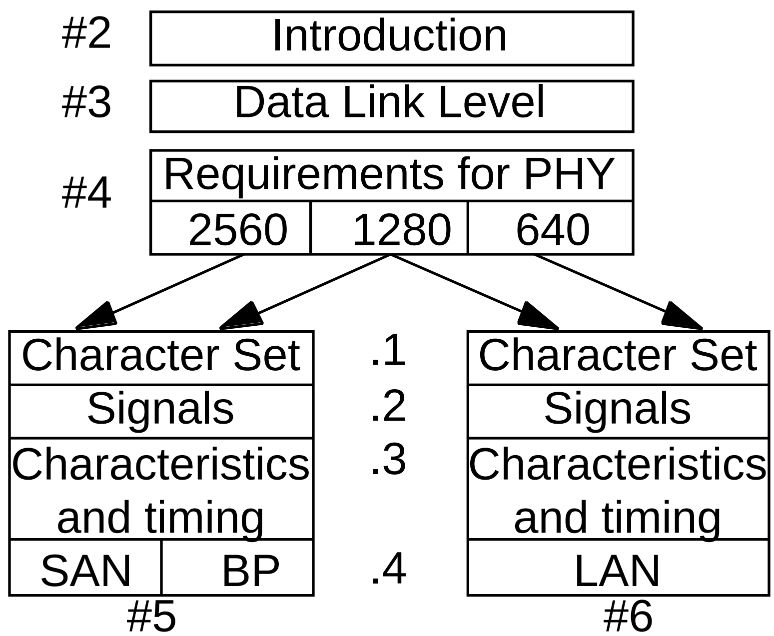


Figure 2.5—1: The map of this document

2.6 General Description of Myrinet *(not a part of the standard)*

Speed

Myrinet connects computing nodes (cards, workstations, and PCs) through full-duplex point-to-point Gigabit-per-second links (e.g., 1.28+1.28 Gbits/sec), and low-latency cut-through switches.

Robustness and Convenience

Any network topology is allowed, because the network is self-configuring. The Myrinet host interfaces may periodically map the network, both to provide fault-tolerance through the dynamic use of alternate routes to circumvent faults, and also as a convenience for installation.

Network Throughput

As a switched network, analogous in structure to Ethernet segments connected by switches, a Myrinet can carry many packets concurrently, each traversing the network at 1.28 Gbits/s. Unlike unswitched Ethernet or FDDI networks, which share a common communication medium, the aggregate traffic capacity of a Myrinet increases with the number of hosts. For example, a Myrinet connecting 16 hosts with a single 16-port switch can carry 16 packets at once, an aggregate traffic capacity (and switch-bisection data rate) of 20.48 Gbits/s.

Versatility

Myrinet packets may be of any length, and thus can encapsulate other types of packets, including IP and special APIs packets, without an adaptation layer.

Each packet is identified by type, so that Myrinet, like Ethernet, can carry packets of many types or protocols simultaneously.

Software Interfaces and End-to-End Performance

Myrinet users achieve short-message latencies between UNIX user processes smaller than 5 μseconds, better than most distributed-memory MPPs (Massively Parallel Processors, or multicomputers, such as the Intel Paragon or the IBM SP-2/3). Users achieve also sustained, one-way data rates (at the user level) exceeding 1 Gbits/s. There exist direct implementations over Myrinet of many of the standard, distributed-memory-MPP software interfaces, including MPI and PVM. There are also Myrinet implementations for a wide variety of operating systems including NT, the many “dialects” of BSD , Linux, Solaris, OSF1, HP-UX, and the realtime operating systems VxWorks.

Technology and Reliability

Myrinet components are implemented with the same advanced technology – full-custom-VLSI CMOS chips – as today's microprocessors that are used in workstations, PCs, and single-board computers. The use of CMOS technology ensures that Myrinet performance will continue to advance in step with advances in the hosts, without changes to the network architecture and software. These CMOS-based Myrinet components are also extremely reliable, having MTBF in the range of several million hours.

[This page intentionally left blank.]

3. Specification of the Data Link Level

This chapter defines the Data Link level specification of Myrinet-on-VME. This level, known as Level 2 of the ISO Reference Model (ISORM) for Open System Interconnection, defines the logical (as opposed to physical) protocol of Myrinet.

3.1 Myrinet Channels, Links, and Ports and Flow Control

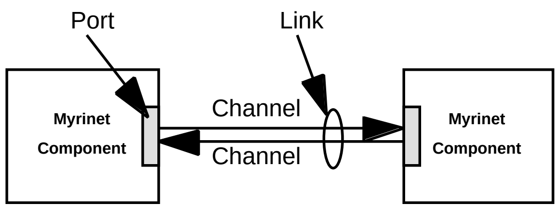


Figure 3.1—1: Myrinet Ports, Channels, and Links

Rule 3.1–1: Myrinet Links

A Myrinet link shall be a full-duplex pair of opposite-going Myrinet channels. A Myrinet channel shall be a unidirectional, point-to-point, communication channel that conveys characters of information. The flow from the sender can be blocked (stopped) temporarily by the receiver at any time, during or between packets, using Flow Control.

Hence, by definition, Myrinet channels appear in links. The existence of the opposite-going channel simplifies automatic network mapping, and could be used at the Physical level to implement flow control and to monitor link continuity.

A Myrinet *Port* is the connection of a link to a component or to a system.

Rule 3.1–2: Character Set

The character set of a Myrinet channel shall include all 256 data bytes and several control symbols.

3.2 Myrinet Flow Control

Rule 3.2–1: Flow Control

The Myrinet Data Link level shall use the data-byte-level flow control, implemented at the Physical level.

Flow control is required for the logical operation of a Myrinet. In addition, it also allows the freedom in the Physical-level implementations for different components and channels within a Myrinet to operate at different data rates.

3.3 Myrinet Packets

Myrinet channels convey packets that are sequences of data bytes. The channel provides the framing of packets, by identifying the first byte (head) of the packet, and the last byte (tail).

Rule 3.3–1: Packet Framing

> The Myrinet Data Link level shall use the packet framing, implemented at the Physical level.

Rule 3.3–2: Packet length (and MTU)

> A Myrinet Data Link level shall convey packets over the channels with any number of data bytes, limited only by an Maximum Transmission Unit (MTU) that shall be at least 4 MBytes.

A byte is the smallest unit of information conveyed on a Myrinet channel.

Myrinet is "byte-wide" at the Data Link level. Independent of whether the Physical-level implementation of a Myrinet channel can carry information as, for example, bit-serial data, byte-parallel data, or multiple bytes in parallel, the Myrinet channel must appear at the Data Link level to be byte-wide. Similarly, software interfaces could restrict data carried from one host to another to multiples of 4 or 8 bytes, but such restrictions appear only at protocol levels higher than the Data Link level.

This specification imposes no limit on the maximum length of a Myrinet packet, but practical system limitations such as timeout[1] periods and buffer sizes limit the packet length.

To be compliant with this Myrinet Data Link specification, the physical layer must support at least packets of 4 MBytes. The software at higher levels can impose more restrictive requirements (e.g., smaller MTU).

Observation 3.3–1: Setting the MTU

> The MTU is not a parameter of Myrinet, but a convention introduced by the application software, above Myrinet. System administrators may set the MTU of their systems to any desired value to meet their special needs. Rule 3.3–2 guarantees that the Myrinet components are capable to handle any MTU of at least 4 MBytes.

---

[1] See section 3.9 (Timeout and Deadlocks)

3.4 Myrinet Packet Format

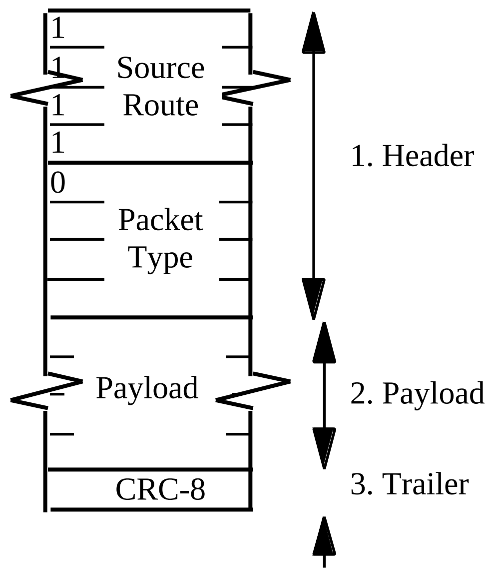


Figure 3.4—1: The Structure of a Myrinet Packet

Rule 3.4–1: Packet Structure

A Myrinet packet shall include in sequence:
1 The header, which shall be four or more bytes.
2 The payload, which shall be zero or more bytes.
3 The trailer, which shall be one byte.

3.4.1 Myrinet Packet Header

Rule 3.4–2: Header Structure

The Myrinet packet header shall include in sequence:
1.1 The source route, which shall be zero or more bytes.
The most-significant bit of each of these bytes shall be 1.
1.2 The packet type, which shall be 4 bytes.
The most-significant bit of the first byte of the packet type shall be 0.

The *source route* is the beginning of a Myrinet packet that is interpreted sequentially by the switches encountered on the route, with each switch "stripping off" (removing) that part of the source route used by that switch. Thus, the source route is sequentially consumed enroute.

Rule 3.4–3: Valid Routing Bytes

For any Myrinet switch, a valid routing byte shall specify an existing port, and also shall have its most significant bit (MSbit) set to 1.

The *packet type* is a field that is used to identify the protocol and suitable software for handling an incoming packet. A Myrinet can carry packets of many types at once, for example, "native" packets used by streamlined application-programming interfaces, and mapping packets used to explore and map the network. The packet type is composed of a two-byte primary type, the most-significant bit of the first of these bytes being 0, followed by a two-byte secondary type. Myricom, Inc., maintains the registry of the primary packet types. Please see the URL: {http://www.myri.com/myri-types.html}. Appendix C lists the assignment of Myrinet packet types, as of July 1998.

Although the header is composed of a variable-length source route followed by a four-byte type , it is always possible to determine from the most significant bits of these bytes where the packet type begins. If all packets followed known routes, this information would be redundant, but mapping packets explore unknown routes. The redundancy in the encoding of the source route in the header of a packet allows a mapping packet routed to a possible host interface to be dropped if it encounters a switch. Similarly, a mapping packet routed to a possible switch can be identified if it encounters a host interface.

3.4.2 Packet Payload

The format and interpretation of the payload part of a Myrinet packet is governed by protocols above the Data Link level. However, it helps in understanding why more information is not required in the Myrinet header to realize that the payload is most often itself a packet that starts with its own header. For example, Myrinet can readily encapsulate Ethernet packets, and thus can carry packets of any protocol supported by Ethernet. The header of the packet carried in the payload can carry information such as the destination address, the packet length, and checksums, which together provide, as required, a degree of error control against data errors or misrouting.

3.4.3 Packet Trailer

Rule 3.4–4: CRC-8

> The last byte of a Myrinet packet shall be originated as the CRC-8 of all the preceding data bytes in that packet, where the CRC-8 is computed using the polynomial X^8+X^2+X+1.

The one-byte trailer of a Myrinet packet contains a cyclic-redundancy-check (CRC-8) character. The CRC-8 detects packets whose data have been corrupted by port-circuit, cable, or connector faults. Packets corrupted by timeout[2] events can additionally be detected at higher protocol levels by the packet being of the wrong length.

The CRC-8 used by Myrinet is the polynomial X^8+X^2+X+1, the same CRC-8 that is used in the header of ATM packets, but without the exclusive-or with hexadecimal 55 (binary 01010101) at the end of the CRC computation. Thus, a packet composed of all zero bytes has a Myrinet-CRC-8 of zero.

[2] See section 3.9 (Timeout and Deadlocks)

Operationally, the CRC-8 is computed on the entire preceding part of the packet, including the header. Because the source-route part of the header is normally modified at each switch, the CRC-8 must be recomputed and checked for each link. At a switch or at an interface, the port circuits compute the CRC-8 of the incoming packet, and substitute the exclusive-or of the computed and received CRC-8 in the trailer of the packet. A received packet whose CRC-8 is correct will have a zero trailer at this point, whereas a packet whose CRC-8 is incorrect will have a non-zero trailer. On the outgoing link, the port circuits compute the CRC-8 as the packet is sent, and exclusive-or the computed CRC-8 into the trailer. If the packet had a correct CRC-8 when received, it will have a correct CRC-8 when sent; whereas if the packet had an incorrect CRC-8 when received, it will have an incorrect CRC-8 when sent.

3.5 Myrinet Components

A Myrinet interface to or within a host computer nominally has one port. A host may include more than one Myrinet interface, but from the network's viewpoint these ports operate independently. The ports of Myrinet interfaces are the only points where new packets are injected into a Myrinet network, and the only points at which they are properly consumed.

A Myrinet switch is a multiple-port component that switches (routes) packets entering on the incoming channel of a port to the outgoing channel of a port selected by a source route in the initial bytes of the packet. A very brief description of Myrinet switching and routing is included in Appendix B.

There is a general rule related to technicalities of "progress" that applies to Myrinet components:

Rule 3.5–1: Progress Guarantee

A Myrinet port shall not block progress indefinitely.

A packet directed to the incoming port of a Myrinet interface must be consumed eventually. Although the interface can block an incoming packet for some interval, for example, while it allocates a new receive buffer, such blocking must be minimized. Similarly, a Myrinet switch must eventually send out on the specified port any packet that it has received, or drop it if that port is blocked for too long (i.e., longer than the timeout[3] period).

Observation 3.5–1: Progress guarantee

The primary mechanism guaranteeing progress and packet consumption is the insistence on consumption at the endpoints of the network, together with the routing being deadlock-free.

If deadlocks occur upon initialization or due to hardware or software errors, the timeout is the fallback mechanism to clear them out.

[3] See section 3.9 (Timeout and Deadlocks)

For any Myrinet switch, a valid routing byte must specify an existing port, and also must have its most significant bit (MSbit) set to 1.

Observation 3.5–2: Dropped Packets

A packet is dropped if it is received and ignored (e.g., not being transmitted by a switch). A dropped packet does not block other packets.

Rule 3.5–2: Invalid Routing

A Myrinet switch shall drop all packets that do not start with a valid routing byte.

Invalid routing bytes can result from software or hardware malfunction, or from unsuccessful explorations by a dynamic mapping process.

A switch drops packets by accepting them without sending them.

3.6 Myrinet Source Routes and switching

Myrinet switching is not specified at the Data Link level. Indeed, the only rule that could be stated is the rule against a rule:

Rule 3.6–1: Packet Routing by Switches

The Data Link-level specification shall not restrict the method of packet routing performed in a Myrinet switch.

Rule 3.6–2: Source Routing

Myrinet switches shall steer packets according to the valid routing bytes at the beginning of the packets, and modify (remove or replace) these bytes.

Observation 3.6–1: Source Routing

By removing routing information from the head of packets, the head is an "iterative" source route.

However, because the design rationale for Myrinet cannot be appreciated without understanding the existing practice in Myrinet routing, Appendix B of this specification offers a brief exposition of that practice.

3.7 Myrinet Unused / Disconnected Ports

.

Rule 3.7–1: Disconnected Ports

> An outgoing packet on a port that is unused, disconnected, or connected through a link to a powered-off component shall not be blocked, but shall be dropped.

A Myrinet must continue to operate under conditions in which some of its ports are unused, some of its links are disconnected, or some of its links are connected to unpowered components. The outgoing packet on a port that is in any of these conditions must not be blocked, or else it would block other links and ports. Instead, outgoing packets on such ports are dropped.

This Data Link level requirement must be implemented at the Physical level. Port circuits can employ signaling such that blocking is disabled on a link that is open or connected to a powered-off component. Alternatively, port circuits can employ methods such as carrier detection or a "heartbeat" protocol to monitor link continuity. In high-availability systems, loss of link continuity or dropped packets can be reported through a monitoring network.

Recommendation 3.7–1: Self-Monitoring

> The Physical layer should continuously monitor the continuity of the links, and issue alerts when discontinuities are discovered.

Sections 5.1 and 6.1 suggest the use of an optional control symbol, Beat, for that purpose.

3.8 Myrinet Topology

The network topology can be viewed as an undirected graph. Any way of linking together interfaces and switches is allowed. The graph can contain cycles; indeed, topologies with cycles are required to provide multiple-path redundancy. The physical network can include unpowered host interfaces and unused switch ports.

3.9 Timeout and Deadlocks

Under certain error conditions a Myrinet is capable of deadlock, a condition in which one or more packets cannot make progress because they are mutually blocked. It is generally the responsibility of the software that establishes sets of routes between hosts to assure that the set of routes is deadlock-free, and it is the responsibility of the hardware to clear any deadlock, if it occurs.

In the event of a deadlock caused by a data error in a packet header or by a software error, it is necessary to clear the deadlock by dropping packets that are contributing to the deadlock.

Rule 3.9–1: Timeout

> If a packet is not terminated for more than a given timeout period since it began, then the port circuits shall terminate and/or drop any such packet.

This rule applies to both sending and receiving ports. Lack of packet termination can result from reasons such as blocking (flow control) or from hardware malfunction.

It is not intended that the timeout be used or depended upon in the normal operation of a Myrinet. Rather, because a Myrinet can be distributed across a large area, this timeout mechanism is regarded as a more practical way to reinitialize a Myrinet than distributing a system-reset signal.

Recommendation 3.9–1: Timeout Period

> The timeout period should be on the order of a second.

Observation 3.9–1: Timeout Period

> This timeout period is expected to be long enough to allow transmission times that manyfold exceeds the MTU to allow for queuing that can occur as a result of "hot spots".

The use of Timeout may significantly vary between installations. Fast-response systems (e.g., military fire control) are expected to require very short timeout periods, whereas non realtime systems may require long periods. Other systems may use yet other mechanisms to monitor progress ignoring the timeout all together.

Suggestion 3.9-1: Setting the Timeout Period

> Designers are encouraged to make it possible to set the timeout period in the field (e.g., by jumpers).

4. The Requirements for the Physical Level

This chapter summarizes the requirements imposed by the Data Link level on the Physical level, and also defines various rates for Myrinet links, regardless of their physical implementations.

A summary of Myrinet requirements from any Physical-level implementation:

- The Physical level channels must convey characters, including all 256 data bytes, and several control symbols.

- The Physical level must provide full-duplex point-to-point links with flow control and packet framing.

- A single byte must be the smallest unit of information conveyed on Myrinet channel. If the physical layer is wider than a byte, padding insertion and removal (if needed) must be provided.

- The MTU of the Physical level must be at least 4 MBytes.

- Packets that require transmission time beyond the timeout must be properly terminated
- (i.e., all bytes are ignored until the next Gap, and a Gap is sent).

- The Physical level must direct packets according to their source route, and modify it.

- The Physical level of a Myrinet channel must provide a CRC-8 packet protection in the packet trailer.

Because all Physical-level implementations must support the above requirements, they differ only in incidental details, not in essential properties. Therefore, it is easy to translate between different Physical implementations using conversion circuits.

4.1 Myrinet Link Rates

Observation 4.1–1: Myrinet-RRR/PHY Designations

> The designation "Myrinet-RRR" specifies any Myrinet implementation that transfers characters at the RRR rate, measured in Mbits/second, regardless of the actual modulation frequency of the links or of the encoding.

The "PHY" defines the specific Physical implementation, such as:

- "SAN" for short cables used on the front and the rear panels, and for PCB traces;
- "BP" for "SAN" over the P0/J0 or P4/J4, used for interconnection with the backplane;
- "LAN" for the longer cables that are used on the front and the rear panels.

4.1.1 640 Mbits/sec Rate

Rule 4.1–1: Nominal Character Period for Myrinet-640

The Nominal character period of Myrinet-640 shall be 12.5 nsec.

4.1.2 1280 Mbits/sec Rate

Rule 4.1–2: Nominal Character Period for Myrinet-1280

The Nominal character period of Myrinet-1280 shall be 6.25 nsec.

4.1.3 2560 Mbits/sec Rate

Rule 4.1–3: Nominal Character Period for Myrinet-2560

The Nominal character period of Myrinet-2560 shall be 3.125 nsec.

These rates are defined uniformly for all the Physical-level implementations (e.g., the same definition of the character period holds both for Myrinet-1280/SAN and for Myrinet-1280/LAN).

Chapter 5 defines Myrinet-1280/SAN, Myrinet-1280/BP, Myrinet-2560/SAN, and Myrinet-2560/BP.

Chapter 6 defines Myrinet-640/LAN and Myrinet-1280/LAN.

5. Specification of the SAN Physical Level

This chapter specifies the SAN (System-Area-Network) Physical-level implementation of Myrinet-on-VME. This level, known as Level 1 of the ISORM, defines the physical (electrical and mechanical, as opposed to logical) protocol of Myrinet.

SAN communication is intended for intra-cluster distances, such as intra-card, intra-subrack, and intra- cabinet. For inter-cluster distances LAN are intended to be used.

This chapter specifies, for SAN Physical implementations, the following:
(5.1) The channel character set (data bytes and control symbols)
(5.2) The encoding of the characters on signals
(5.3) The electrical characteristics of these signals (such as timing and voltage)
(5.4) Pin-assignments, connectors, and cables.

Section (5.4) includes the specifications for front-panel and rear-panel applications using SAN connectors and SAN cables, and also for backplane applications using P0 or P4 connectors.

The designation of the backplane implementation is "BP", which is SAN using the P0/J0 or the P4/J4 connectors. Hence, every specification of SAN (except the details of the SAN connectors) applies also to BP.

5.1 The SAN Channel Character Set and Flow Control

Rule 5.1–1: The SAN character set

The SAN character set shall include, in addition to all 256 data bytes, at least the following control symbols: Gap, Res1, and Res2.

The Gap symbol is used for packet framing. Res1 and Res2 are reserved for future use, such as for interoperability with future versions of Myrinet as they evolve with time.

Rule 5.1–2: Gap Symbol

The Gap symbol shall indicate that the previous data byte was the trailer of the previous packet, and that the next data byte (i.e., not a control symbol) will be the first byte of the header of the next packet.

Rule 5.1–3: Asynchronous Control Symbols

The Gap symbol shall be sent only between packets. All other control symbols can be sent either between or inside packets.

Permission 5.1–1: Repeated Gap Symbols

Successive Gap symbols may be sent repeatedly between packets.

Recommendation 5.1–1: Beat Symbol (Continuity Monitoring)

The Beat symbol should be used to monitor the continuity of links. Beat should be sent at a the frequency defined below. If no Beat arrives within a given period an alert should be generated regarding that link.

Rule 5.1–4: Beat Frequency

If the Beat symbol is used, it shall be sent every 10 microseconds (±10%). The receiver's timeout period is left for implementer's discretion.

The flow control is handled on SAN channels by using the B-bit (the "Block signal"), as described in section 5.2.

Rule 5.1–5: Flow Control

Upon receiving an asserted IB signal, the output channel of a port shall stop its transmission of data bytes until the IB is unasserted.

Observation 5.1–1: Flow Control Applicability

The flow control applies only to data bytes, not to control symbols.

Observation 5.1–2: Bytes in Flight

From the time a component issues a Block signal (by asserting its OB) to block the flow of incoming bytes, until the flow actually stops, more bytes may still arrive. The number of these bytes is N=2LR/c'+K, where L is the length of cable, R the transmission rate, c' the speed of electronic propagation over this cable (typically c'=0.6c=180Mm/s in copper), and K is the number of bytes transmitted since the IB arrives until the flow actually stops.

For example, a copper cable of L=3m, at the rate of R=160MB/s, with K=5B can still deliver after a Block signal is asserted, at least:

N = 2*3m*160M(B/sec) / (180Mm/sec) + 5B = 11B.

Rule 5.1–6: Cable Slack Buffers to Prevent Data Loss

Components that receive data from cables shall have enough slack-buffer memory to absorb the data that is already in flight when a Block signal is issued, thus preventing data loss. This amount compensates for the cable length and for the time it takes senders to block the flow after receiving the Block signal.

Recommendation 5.1–2: Cable Slack Buffers to Prevent Data Starvation

In order to prevent potential "data starvation" by receivers after unblocking the data flow, the size of the slack buffers should be at least twice of what is needed just to prevent data starvation.

5.2 The Encoding of the SAN Characters on Signals

The SAN physical layer for a Myrinet link is comprised of 20 signals organized in two independent channels of 10 signals for communication in each direction.

Rule 5.2–1: SAN signals

The organization for each SAN channel is:
8 data bits (0 through 7)
1 Data/Control bit (D: 1 = Data, 0 = Control)
1 Flow Control "Block" bit (B: 1 = stop, 0 = go)

Rule 5.2–2: Encoding of the SAN Character Set

The encoding of the SAN character set shall be as shown in Table 5–1.

	Bit:	D	7	6	5	4	3	2	1	0
Mandatory	Data Byte	1	d	d	d	d	d	d	d	d
Mandatory	Gap Symbol	0	0	0	0	0	1	1	0	0
Mandatory	Res1 Symbol	0	0	0	0	0	0	0	1	1
Mandatory	Res2 Symbol	0	0	0	0	0	1	1	1	1
Optional	Beat Symbol	0	1	1	1	1	1	1	1	1

Table 5–1: Encoding of SAN Characters

The signals are denoted by either "I*" (inputs from the Myrinet into the component) or "O*" (outputs to the Myrinet from the component) for * being any of {0,1,2,3,4,5,6,7,D,B}. The SAN cables connect each O* to the corresponding I*. Hence, the O* of a port is the I* of the port with which it is connected.

The direction of B, the flow-control signal, is opposite to that of all other signals. It is an output of the component for which the rest of the channel is an input.

Hence, the output channel consists of {O0,O1,O2,O3,O4,O5,O6,O7,OD,IB} and the input channel consists of {I0,I1,I2,I3,I4,I5,I6,I7,ID,OB}.

OB and IB are used for flow control. IB is an input to the sending part of a port, and commands the port to stop sending. Similarly, OB is an output from the receiving part of a port, and signals the sending part of the port on the other side of the link to stop sending.

Recommendation 5.2–1: Unknown Characters

Characters that are not defined by Rule 5.2–2 should be ignored or treated as errors.

Rule 5.2–3: SAN NRZI Encoding

The 8 data signals and the D signal shall be NRZI (Non-Return-to-Zero, Invert, or "transition") encoded, such that a transition on a line represents a 1 and no transition represents a 0.

The B signal is level encoded (not NRZI) with HIGH being blockin, and LOW being non-blocking.

A Myrinet receiver operates asynchronously, and accepts characters whenever they arrive; however, they cannot be separated in time by less than the character period.

All characters include at least one transition. Myrinet receivers group together the transitions of multiple signals within a certain amount of time ("window") after the first transition is detected on any signal, and treat all of them together as a single character.

Observation 5.2–1: Signal frequency

This encoding keeps the maximum fundamental frequency on any line from exceeding half of the data rate. For example, while transferring data at 160Mbits/s per signal line, no signal exceeds 80MHz in the fundamental frequency.

Rule 5.2–4: Skew

The total skew (difference in arrival time from earliest to latest transition) from sender to receiver shall be less than 40% the character period.

Recommendation 5.2–2: Skew

To maximize timing margins and improve bit error rate, the total skew of the PCB traces, connectors, and cable should be made as small as possible.

Permission 5.2–1: Skew reduction

Adaptive techniques may be used to reduce the effective skew at the receiving port.

Recommendation 5.2–3: Cable Skew

The skew of any cable should not exceed one quarter of the character period.

5.3 The Characteristics of SAN Signals

5.3.1 Packet Timing

Rule 5.3–1: IPG (Inter-Packet Gap)

Myrinet receivers shall be able to handle packets with an Inter-Packet Gap of (minimum) one Gap symbol.

5.3.2 SAN Signal Characteristics

Rule 5.3–2: SAN Signal characteristics

The SAN signals shall have controlled edge rate and shall have the characteristic specified in Table 5–2.

	Level	Character Period	Switching time Min	Switching time Max	Source Terminatio n	Type
Myrinet-1280	1.25 V	6.250 ns	0.5 ns	2.0 ns	50 Ω	Single Ended
Myrinet-2560	1.25 V	3.125 ns	0.2 ns	1.0 ns	50 Ω	Single Ended

Table 5–2: SAN Signaling

Recommendation 5.3–1: Pulldown for Input Signals

All SAN input signals (I*) should have high impedance pulldown of at least 5 KΩ (20 KΩ nominal).

These pulldowns maintain floating inputs at a voltage sufficiently far from the switching threshold to avoid noise inputs from unused ports or disconnected cables.

Rule 5.3–3: Common GND for SAN Signaling

Myrinet components that use SAN links shall have common GND.

Recommendation 5.3–2: Common GND for SAN Signaling

The best way to achieve common ground is by using the same power+GND source. If this is not the case, ground straps should be used, but this still does not guarantee common GND, e.g., between "floating chasses".

5.4 SAN Connectors, Cables, and Pin-Assignments

This section specifies low-level details of the connectors, and recommends certain connectors and cables. SAN connection through the front and/or the rear panels must be as specified in section 5.4.1, whereas SAN connection through the backplane must use either the P0 or the P4 connectors, as specified in section 5.4.2.

Rule 5.4–1: Connector Gender

Myrinet components shall have receptacle (female) connectors. Myrinet cables shall have plug (male) connectors.

5.4.1 Front-Panel and/or Rear-Panel Connection (SAN)

This section describes the SAN connection scheme. Items covered are the SAN connectors and their locations, and the basic SAN pinout.

Rule 5.4–2: SAN Connectors

For applications when the SAN connector are provided on the front panel of the VME board, the board mounted SAN connector should be a 40-pin Microstrip receptacle [4].

Recommendation 5.4–1: Positioning of SAN Connectors.

The SAN connector should be positioned as close as possible to the board edge and the front panel should be designed to allow for the latching cable connector.

Observation 5.4–1: Dual Link SAN Connectors

Since the SAN connector has 40 pins, and since a SAN link uses only 20 pins, each SAN connector carries 2 full-duplex links, designated as A and B. The A link uses the 20 middle pins (with the output pins 11 to 20 of each device connected to the input pins 30 to 21, respectively, of the other device; and the B link uses the 20 outside pins (with the output pins 1 to 10 connected to the input pins 40 to 31).

Rule 5.4–3: The A Link is the Primary Link

The A link shall be considered as the primary link of the connection. Components that use one link only shall use the A link.

[4] The board mounted connector should be the 40-Pin right-angle Microstrip receptacle, AMP part number 536295-1, or equivalent.

Rule 5.4–4: Unused SAN Links

When only one link is used by Myrinet, the other link (on dual link connectors) shall not be used for other purposes.

Rule 5.4–5: SAN Pin Assignment

The pin assignments in the board mounted SAN connector shall be as shown in Table 5–3. The signals S* (for * being any of 0,1,2,3,4,5,6,7,D,B) are the output signals of the B link, similar to O* of the A link. The R* are the inputs of the B link, similar to the I* of the A link.

Sending Channels			Receiving Channels		
Pin #	Signal Name	Link	Link	Signal Name	Pin #
1	S0	B	B	R0	40
2	S1	B	B	R1	39
3	S2	B	B	R2	38
4	S3	B	B	R3	37
5	S4	B	B	R4	36
6	S5	B	B	R5	35
7	S6	B	B	R6	34
8	S7	B	B	R7	33
9	SD	B	B	RD	32
10	SB	B	B	RB	31
11	O0	A	A	I0	30
12	O1	A	A	I1	29
13	O2	A	A	I2	28
14	O3	A	A	I3	27
15	O4	A	A	I4	26
16	O5	A	A	I5	25
17	O6	A	A	I6	24
18	O7	A	A	I7	23
19	OD	A	A	ID	22
20	OB	A	A	IB	21

Table 5–3: Dual Link SAN Connector Pinout

5.4.1.1 SAN Cables

Rule 5.4–6: SAN Cables

The SAN cables shall have 40-pin Microstrip plugs[5] that mate with the board mounted receptacles, and shall connect the signals defined in Rule 5.4–5 such that they connect each Ox signal of one component with the Ix of the other component, and each Sx to the other Rx.

Observation 5.4–2: SAN Cable Connectivity

Each connector pin#N is connected with connector pin#(41-N) at the other cable end (e.g., pin 1 of each connector to pin 40 of the other connector, pin 2 to pin 39, pin 3 to pin 38 and so on).

Each row of Table 5–3 indicates the pins (of the opposite ends) that are interconnected by the SAN cable.

Rule 5.4–7: SAN Cables for Myrinet-1280

Myrinet-1280/SAN shall be used only over SAN cables not longer than 10ft.

Rule 5.4–8: SAN Cables for Myrinet-2560

Myrinet-2560/SAN shall be used only over SAN cables not longer than 10ft.

Rule 5.4–9: SAN Cables – Impedance

The SAN cables shall have 50 Ω controlled impedance.

5.4.1.2 VME Front and/or Rear-Panel Access

A SAN cable connects to the SAN connector on the VME board through the Front Panel or boards that are on the bulkhead. The board-mounted connector is specified in Rule 5.4–6. Since connectivity is through a cable, the connector on the boards can be located anywhere on the Front Plate that the board designer chooses.

Permission 5.4–1: Right-side-up Connectors

The connector may be placed "right side up" for VME board mounting.

[5] The SAN cables should be 40-signal Microstrip cables, AMP part number 636349 (or equivalent).

Permission 5.4–2: Right-side-down Connectors

The connector may be placed "upside down" for mounting on a mezzanine board such as a PMC.

Permission 5.4–3: SAN Connector Location

The SAN connector may be placed any place along the edge of the VME board.

5.4.2 Backplane Connection (P0/J0 and/or P4/J4)

This section specifies the Myrinet connection through the backplane, using either P0/J0 or P4/J4.

Observation 5.4–3: Myrinet-RRR/BP

The Backplane Connection (via P0/J0 and/or P4/J4) has the Physical-level designation of "BP". Except for the mechanical details of the connectors Myrinet-RRR/BP is identical to Myrinet-RRR/SAN.

Rule 5.4–10: Conformity with VME64x

The P0/J0 connector shall conform to ANSI/VITA 1.1-1997, VME64x, chapter 4, and the P4/J4 connector shall conform to ANSI/VITA 1.3-1997, VME64x 9U x 400 mm Format, chapter 4.

Observation 5.4–4: Caution on P0 and P4 Connectors

VME-6U and VME64x-9U boards using a P0 connector (not specified in this document) could conflict with VME backplanes that have a mechanical structure member between J1 and J2 connectors.

Similarly, VME64x-9U boards using a P4 connector (not specified in this document) could conflict with VME backplanes that have a mechanical structure member between J2 and J3 (or between J2 and J5) connectors.

Rule 5.4–11: Pin Assignment for the Backplane Connectors

The pin assignments for the P0 connector shall be as shown in Table 5–4.

Pos #	Row A	Row B	Row C	Row D	Row E
1	Res	GND	Res	GND	Res
2	GND	S0	GND	S1	GND
3	S2	GND	S3	GND	S4
4	GND	S5	GND	S6	GND
5	S7	GND	SD	GND	SB
6	GND	O0	GND	O1	GND
7	O2	GND	O3	GND	O4
8	GND	O5	GND	O6	GND
9	O7	GND	OD	GND	OB
10	GND	Res	GND	Res	GND
11	IB	GND	ID	GND	I7
12	GND	I6	GND	I5	GND
13	I4	GND	I3	GND	I2
14	GND	I1	GND	I0	GND
15	RB	GND	RD	GND	R7
16	GND	R6	GND	R5	GND
17	R4	GND	R3	GND	R2
18	GND	R1	GND	R0	GND
19	Res	GND	Res	GND	Res

Table 5–4: Backplane Connector Pinout

The signals S* (for * being any of 0,1,2,3,4,5,6,7,D,B) are the output signals of the B link, similar to O* of the A link. The R* are the inputs of the B link, similar to the I* of the A link. The “Res” indicate a reserved signal for future use.

Observation 5.4–5: P0 has 7 rows (copied from VME64x Observation 4.6)

The J0/RJ0 connectors have seven physical rows of contacts, with the z and f contact rows connected to the backplane's ground plane. On the VME64x board, there is no z row of contact holes. Depending on the connector design, the P0 and the RP0 connector ground contacts in the z and f row of the connector (which is on the connector shroud) will alternately connect to the board's f row of ground contacts.

6. Specification of the LAN Physical Level

This chapter specifies the LAN (Local-Area-Network) Physical-level implementation of Myrinet-on-VME. This level, known as Level 1 of the ISORM, defines the physical (electrical and mechanical, as opposed to logical) protocol of Myrinet.

LAN communication is intended for distances of up to a few tens of meters. For distances up to 3 meter (typical for intra-cabinet and between colocated cabinets) SAN can be used.

This chapter specifies, for LAN Physical implementations, the following:
(6.1) The channel character set (data bytes and control symbols)
(6.2) The encoding of the characters on signals
(6.3) The electrical characteristics of these signals (such as timing and voltage)
(6.4) Pin-assignments, connectors, and cables.

Section (6.4) includes the specifications for front-panel and rear-panel applications using LAN connectors and LAN cables. Unlike SAN, LAN cannot be used for backplane applications that use P0 or P4 connectors.

6.1 The LAN Channel Character Set and Flow Control

Rule 6.1–1: The LAN character set

> The LAN character set shall include (in addition to all 256 data bytes) at least the following control symbols: Gap, Stop, and Go.

The Gap symbol is used for packet framing. The Stop and Go symbols are used for flow control.

Rule 6.1–2: Gap Symbol

> The Gap symbol shall indicate that the previous data byte was the trailer of the previous packet, and that the next data byte will be the first byte of the header of the next packet.

Rule 6.1–3: Asynchronous Control Symbols

> The Gap symbol shall be sent only between packets. All other control symbols can be sent either between or inside packets.

Permission 6.1–1: Repeated Gap Symbols

> Successive Gap symbols may be sent repeatedly between packets.

Recommendation 6.1–1: Beat Symbol (Continuity Monitoring)

The Beat symbol should be used to monitor the continuity of links. Beat should be sent at a the frequency defined below. If no Beat arrives within a given period an alert should be generated regarding that link

Rule 6.1–4: Beat Frequency

If the Beat symbol is used, it shall be sent every 10 microseconds (±10%). The receiver’s timeout period is left for implementer’s discretion.

Rule 6.1–5: Flow Control (Stop and Go Symbols)

Upon receiving the Stop symbol, the output channel of a port shall stop its transmission of data bytes until a Go symbol is received.

Observation 6.1–1: Flow Control on Opposite Channel

The flow-control symbols that control the flow from a source to a destination are inserted in the opposite-going channel, of that link, from the destination to the source.

Observation 6.1–2: Flow Control Applicability

The flow control applies only to data bytes, not to control symbols.

Permission 6.1–2: Repeated Flow Control Symbols

Successive Flow Control symbols (Stop or Go, whichever is applicable at the time) may be sent repeatedly when data bytes are stopped, and between packets.

Observation 6.1–3: Bytes in Flight

From the time a component sends the Stop symbol to block the flow of incoming bytes until the flow actually stops, more bytes can still arrive. The number of these bytes is N=2LR/c’+K, where L is the length of cable, R the transmission rate, c’ the speed of electronic propagation over this cable (typically c’=0.6c=180Mm/s in copper), and K is the number of bytes transmitted since the Stop signal arrives until the flow actually stops.

For example, a copper cable of L=10m, at the rate of R=160MB/s, with K=5B can still deliver after a Stop symbol is sent, at least:

N = 2*10m*160M(B/sec) / (180Mm/sec) + 5B = 23B.

Rule 6.1–6: Cable Slack Buffers to Prevent Data Loss

Components that receive data from cables should have enough slack-buffer memory to absorb the data that is already in flight when a Block signal is asserted. This amount should compensate for the cable length and for the time it takes senders to block the flow after receiving the Stop symbol. This amount of slack buffers is needed to prevent data loss.

Recommendation 6.1–2: Cable Slack Buffers to Prevent Data Starvation

In order to prevent potential "data starvation" by receivers after unblocking the data flow, the size of the slack buffers should be at least twice of what is needed just to prevent data starvation.

6.2 The Encoding of the LAN Characters on Signals

The LAN physical layer for a Myrinet link is comprised of 18 signals organized in two independent channels of 9 signals for communication in each direction.

Rule 6.2–1: LAN signals

The organization for each LAN channel is:
8 data bits (0 through 7)
1 Data/Control bit (D: 1 = Data, 0 = Control)

Rule 6.2–2: Encoding of the LAN Character set

The encoding of the LAN character set shall be as shown in Table 6–1.

	Bit:	D	7	6	5	4	3	2	1	0
Mandatory	Data Byte	1	d	d	d	d	d	d	d	d
Mandatory	Gap Symbol	0	0	0	0	0	1	1	0	0
Mandatory	Go Symbol	0	0	0	0	0	0	0	1	1
Mandatory	Stop Symbol	0	0	0	0	0	1	1	1	1
Optional	Beat Symbol	0	1	1	1	1	1	1	1	1

Table 6–1: Encoding of LAN Characters

The signals are denoted by either "I*" (inputs from the Myrinet into the component) or "O*" (outputs to the Myrinet from the component) where * being any of {0,1,2,3,4,5,6,7,D}. The LAN cables connect each O* to the corresponding I* at the other end. Hence, the O* of a port, is connected with the I* of the port at the other end.

Hence, the output channel consists of {O0,O1,O2,O3,O4,O5,O6,O7,OD} and the input channel consists of {I0,I1,I2,I3,I4,I5,I6,I7,ID}.

Recommendation 6.2–1: Unknown Characters

> Characters that are not defined by Rule 6.2–2 should be ignored or treated as errors.

Rule 6.2–3: LAN Differential Signaling

> The 8 data signals and the D signal shall be differentially transmitted over a twisted-pair of wires.

Rule 6.2–4: NRZI Signaling

> All the LAN signals shall be NRZI (Non-Return-to-Zero, Invert, or "transition") encoded, such that a transition on a line represents a 1 and no transition represents a 0.

A Myrinet receiver operates asynchronously, and accepts characters whenever they arrive; however, they cannot be separated in time by less than the character period.

All characters include at least one transition. Myrinet receivers group together the transitions of multiple signals within a certain amount of time ("window") after the first transition is detected on any signal, and treat all of them together as a single character.

Observation 6.2–1: Signal frequency

> This encoding keeps the maximum frequency on any line from exceeding half of the data rate. For example, while transferring data at 160Mbit/s per line, no signal exceeds 80MHz, in the fundamental frequency.

Rule 6.2–5: Skew

> The total skew (difference in arrival time from earliest to latest transition) from sender to receiver shall be less than 40% the character period.

Recommendation 6.2–2: Skew

> To maximize timing margins and improve bit error rate, the total skew of the PCB traces, connectors, and cable should be made as small as possible.

Permission 6.2–1: Skew reduction

> Adaptive techniques may be used to reduce the effective skew at the receiving port.

Recommendation 6.2–3: Cable Skew

> The skew of any cable should not exceed one quarter of the character period.

6.3 The Characteristics of LAN Signals

6.3.1 Packet Timing

Rule 6.3–1: IPG (Inter-Packet Gap)

Myrinet receivers shall be able to handle packets with an Inter-Packet Gap of one Gap symbol only.

6.3.2 LAN Signal Characteristics

Rule 6.3–2: LAN Signal characteristics

The LAN signals shall have controlled edge rate and shall have the characteristic specified in Table 6–2.

	Level	Character Period	Switching time Min	Switching time Max	Source Termination	Type
Myrinet-640	1.2+/-0.4 V	12.50 ns	0.8 ns	4.0 ns	100 Ω	differential
Myrinet-1280	1.2+/-0.4 V	6.25 ns	0.5 ns	2.0 ns	100 Ω	differential

Table 6–2: LAN Signaling

Recommendation 6.3–1: Pulldown for Input Signals

All LAN input signals (I*) should have high impedance pulldown of at least 10KΩ.

Rule 6.3–3: Common GND for LAN Signaling

Myrinet components that use LAN links shall have GND levels that agree within the common-mode range.

6.4 LAN Connectors, Cables, and Pin-Assignments

This section specifies low-level details of the connectors, and recommends certain connectors and cables. LAN is used only to enter/egress subracks on the front or on the rear panels.

Rule 6.4–1: Connector Gender

Myrinet components shall have receptacle (female) connectors. Myrinet cables shall have plug (male) connectors.

6.4.1 Front-Panel and/or Rear-Panel Connectors (LAN)

Rule 6.4–2: LAN Connectors

The standard cable-connector for Myrinet ports shall be a 37-pin D-Subminiature Connector (DB-37). The component-mounted connectors shall be receptacle (female), and the cable-end connectors shall be plug (male).

In addition to these 37 pins, the cable's shield with its shell shall be connected to GND at both ends. When unshielded cables (e.g., "twist-n-flat" ribbon cables) are used, at least one wire shall carry GND between the end shells.

Rule 6.4–3: LAN Pin Assignment

The pin assignments in the connectors shall be as shown in Table 6–3.

A		B		C		D	
pin#	Name	pin#	Name	pin#	Name	pin#	Name
1	Rd+	19	Sd+	20	Rd-	37	Sd-
2	S0+	18	R0+	21	S0-	36	R0-
3	S1+	17	R1+	22	S1-	35	R1-
4	R7+	16	S7+	23	R7-	34	S7-
5	R6+	15	s6+	24	R6-	33	S6-
6	S2+	14	R2+	25	S2-	32	R2-
7	S3+	13	R3+	26	S3-	31	R3-
8	R5+	12	S5+	27	R5-	30	S5-
9	R4+	11	S4+	28	R4-	29	S4-
The cable's Pin#10 must be grounded if it is longer than 10m							

Table 6–3: LAN Connector Pinout

Each signal is encoded differentially between pins that are in the same row, either of columns A–C or of columns B–D.

Rule 6.4–4: LAN Cable Connectivity

The LAN cable shall connect each Sx to its corresponding Rx, as defined in Rule 6.4–3. This connects each pin from column A with the same-row pin from column B (i.e., pin 1 with pin 19, 2 with 18,..., and 9 to 11), and similarly pins from column C with same-row pins from column D (i.e., 20 with 37, 21 with 36, ..., and 28 with 29).

Recommendation 6.4–1: Twisted Pairs

The LAN cables should consist of at least 18 twisted pairs, a single wire for pin#10, and an electrical connection of the shells (by a cable shield or by a wire) .

Permission 6.4–1: LAN Connector Location

The LAN connector may be placed any place along the front panel of the VME board.

Rule 6.4–5: LAN Cables

The LAN cables shall have connectors that mate with the connector specified in Rule 6.4–2 such that they connect each Sx signal with its corresponding Rx at the other end. The LAN cables shall have 100Ω controlled impedance.

Links between Myrinet components could employ a variety of cable types in the 100Ω to 110 Ω range of characteristic impedance. Depending upon the EMI requirements, either unshielded or shielded cable may be used.

Rule 6.4–6: Pin #10 of the DB-37 Connector

LAN cables that are longer than 10 meters shall have pin#10 (on their connectors), connected to GND at both ends.

Rule 6.4–7: LAN Cables for Myrinet-640

Myrinet-640/LAN shall be used only over LAN cables that are not longer than 25 meters, and have pin #10 connected to the shell/shield GND.

Rule 6.4–8: LAN Cables for Myrinet-1280

Myrinet-1280/LAN shall be used only over LAN cables that are not longer than 10 meters, and have pin#10 floating (i.e., not connected to anything).

Rule 6.4–9: Auto-Speed for LAN ports

A Myrinet port shall send at the Myrinet-640 rate if pin#10 of the LAN cable is connected to GND, and shall send at the Myrinet-1280 rate if it is not connected to GND (indicating that the cable is not longer than 10 meters).

[This page intentionally left blank.]

7. Appendix A: Examples of Myrinet-on-VME

7.1 Examples

The following figure demonstrates 4 SBCs interconnected over the front panel with a switching board (shown in the middle) providing internal and external connectivity.

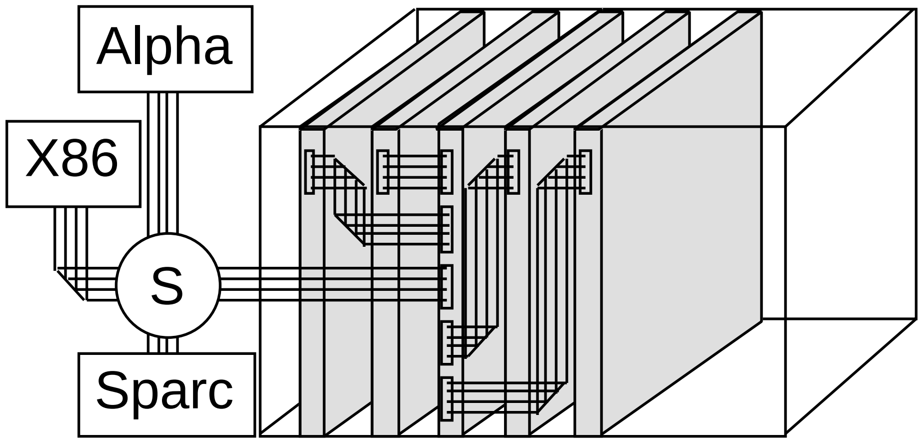


Figure 7.1—1: Front Panel Interconnection of 4 SBCs

Following is an example of a backplane overlay that connects 4 P0 connectors (with a total of 8 links) to 4 SAN connectors, using two 8x8 switches to provide both internal communication and flexible communication with the outside, in several possibilities ranging from using four SAN cables to using one SAN cable for accessing the four boards (that are behind these P0 connectors).

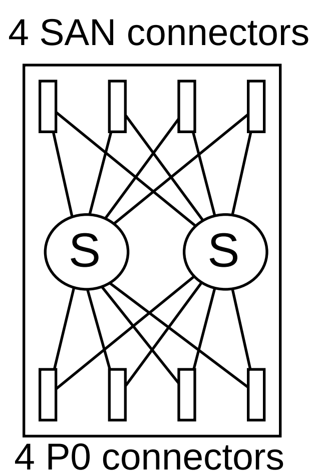


Figure 7.1—2: A Backplane Overlay (with Myrinet/BP and Myrinet/SAN)

The following figure shows how 4 such overlays may be used for interconnecting 16 boards with the full bisection bandwidth (of 10.28Gigabit/sec for Myrinet-1280/BP).

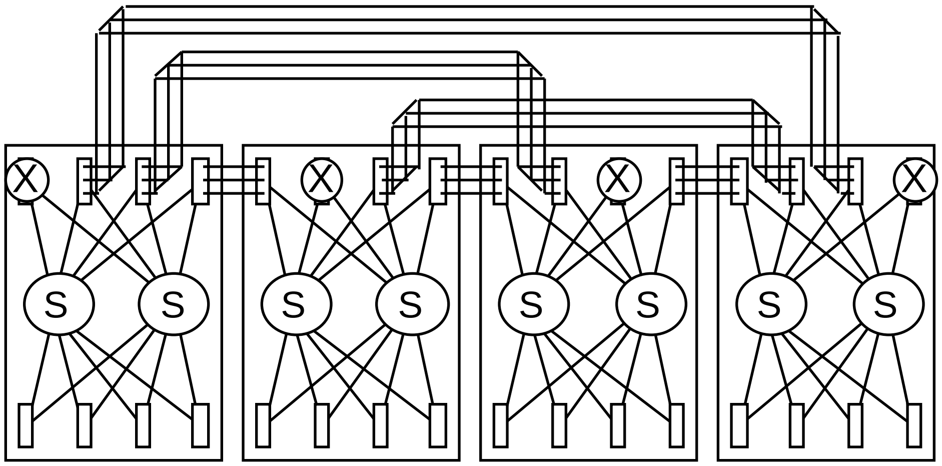


Figure 7.1—3: Full-Bisection-Bandwidth 16-Board Configuration

The following figure shows a subrack with 16 processing boards, each with an on-board Myrinet connecting 4 processing nodes (such as microprocessors and FPGA nodes) forming together a 64-processor system in a subrack.

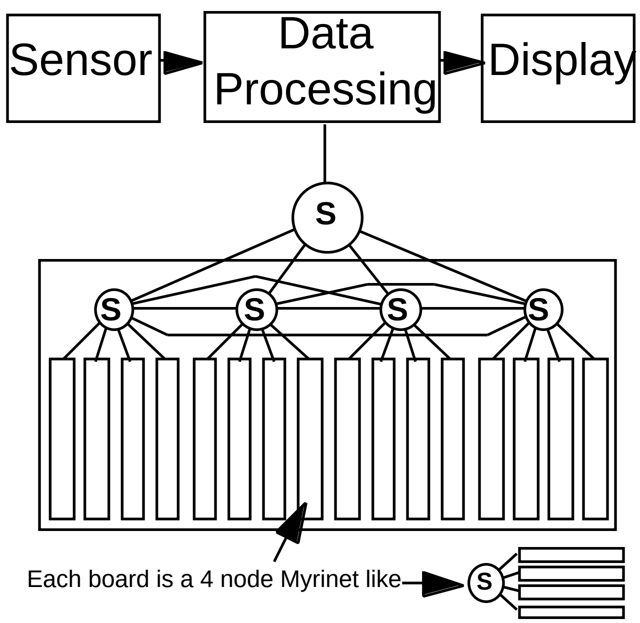


Figure 7.1—4: 64 Processors in a Subrack

8. Appendix B: Existing Practice in Myrinet Routing
(Informative - not a part of the standard)

This Appendix is not a part of the Myrinet specification. It offers a brief exposition of the existing practice in Myrinet routing, in order to help designers familiarize themselves with the Myrinet technology and with the rationale behind it.

Cut-Through Routing

Myrinet switches may employ any type of routing, including store-and-forward routing, but Myrinet was designed to exploit the low latency of cut-through routing. In cut-through routing, the incoming packet is advanced into the selected outgoing channel immediately, provided that the selected outgoing channel is not already occupied by another packet. The packet is then spooled through this established path until the path is broken by the tail of the packet. If the selected outgoing channel is occupied by another packet, the incoming packet is blocked.

Source-Route Encoding

The source-route part of the packet header is generated initially by the hosts or host interfaces, and is interpreted by the switches. The encoding of this source route is at the option of the designers and manufacturers of Myrinet switches.

The Myrinet switches employ cut-through routing based on source routes encoded in single bytes, each with a most-significant bit of 1. Each switch uses a single source route byte and then removes it from the packet. It is expected that future switches of degree higher than 64 ports will use and remove multiple bytes of source route.

Existing Myrinet switches interpret the leading source-route byte as the difference between the outgoing port number and the incoming port number. The advantage of relative port addressing over absolute port addressing is that it is then possible, without knowing the complete map of a network, to "reverse" a route, e.g., for a route from host A to host B of {+3, -2, +5}, the reverse route from host B to host A is {-5, +2, -3}.

Myrinet switches drop packets whose first byte has a most significant bit of 0.

The computation of the outgoing port number does not "wrap," e.g., in a 16-port switch a packet arriving on port 2, with a relative port of -3 at the head of the packet, yields an out-port of -1, which is not valid (and is not wrapped to mean port 15).

Packets addressed to non-existing ports (e.g., to port 19 of a 16-port switch) are dropped. Hence, the valid relative-ports for an 8-port switch, for example, are as listed in the table below.

Extensions to the operation of Myrinet switches could be accessed or controlled through packets addressed to invalid switch ports.

To:	Port 0	Port 1	Port 2	Port 3	Port 4	Port 5	Port 6	Port 7
From Port 0	0	1	2	3	4	5	6	7
From Port 1	-1	0	1	2	3	4	5	6
From Port 2	-2	-1	0	1	2	3	4	5
From Port 3	-3	-2	-1	0	1	2	3	4
From Port 4	-4	-3	-2	-1	0	1	2	3
From Port 5	-5	-4	-3	-2	-1	0	1	2
From Port 6	-6	-5	-4	-3	-2	-1	0	1
From Port 7	-7	-6	-5	-4	-3	-2	-1	0

Table 8–1: The Relative Port Routing for a Myrinet 8-Port Switch

9. Appendix C: Assignment of Myrinet Packet Types

Assignment of Myrinet Types (as of July-31-1998)

(Copied from URL="http://www.myri.com/myri-types.html")

Type (range)	Assigned for (assigned to)
0x0000	Reserved (Myricom)
0x0001	V2 Data Packet (Myricom)
0x0002	V2 Mapping Packet (Myricom)
0x0003	V2 Data Packet (Myricom)
0x0004	Data Packet (Myricom)
0x0005	MyriMap (Myricom)
0x0006	MyriProbe (Myricom)
0x0007	MyriOption (Myricom)
0x0008	GM (Myricom)
0x0009-0x000E	Reserved (Myricom)
0x000F	Mapping (Myricom)
0x0010	Link test packets (Myricom)
0x0011-0x001F	Reserved (Myricom)
0x0020	Ethernet over Myrinet (GM) (Myricom)
0x0021-0x00FF	Reserved (Myricom)
0x0100	Active Messages (UC Berkeley)
0x0110	SSN-Multicasting (UCLA CS Dept.)
0x0200-0x0207	Fast Messages (FM)
0x0300	PacketWay Protocol (PacketWay)
0x0301-0x0302	BDM MCP (MPICH, Mississippi State U)
0x0310	Reserved (Myricom)
0x0330	CSPI packets (CSPI)
0x0340	MPI (MPI Software Technology)
0x0350	BIP (LHPC)
0x0360	Reserved (Myricom)
0x0370	CLF (Digital Equipment Corporation)
0x0400	PM (Real World Computing Partnership)
0x0500	MPI (Hughes Aircraft Co., MPI)
0x0600-0x0601	Isotach (UVa CS Dept. Isotach)
0x0700	VIA: Virtual Interface Architecture (Intel)
0x0800	Portals (Sandia)
0x0900	U-Net (Cornell)
0x0A00	Sanders Packets (Sanders)
0x0B00	AFRL Packets (Air Force Research Lab (AFRL))
0x1000-0xFFFF	Reserved (Myricom)

Consult the above URL for the current assignment of Myrinet packet types.

Requests for Myrinet packet types should be e-mailed to <help@myri.com>.

[This page intentionally left blank.]

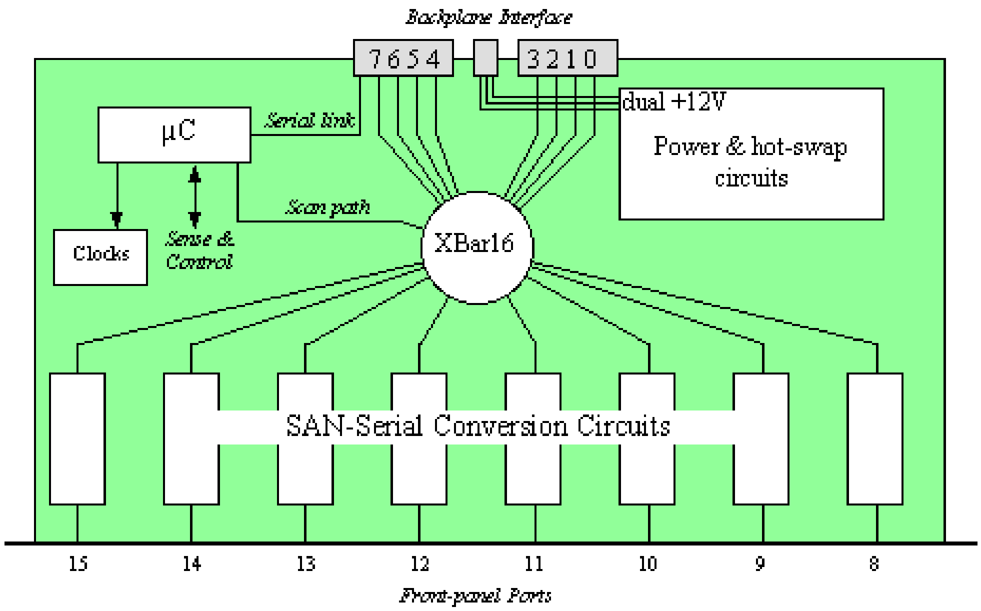
10. Appendix D: Glossary

BP	Backplane
Character	9 signals conveying a data byte or a control symbols
CRC	Cyclic-Redundancy-Check
Channel	A one-way, point-to-point, byte-wide communication medium
GND	Electrical "ground"
Gbits	Giga bits
Hot Spot	Congestion of packet traffic
I0-I7	Primary port (A-link) input data bits
IB	Primary port (A-link) input Blocking bit
ID	Primary port (A-link) input Data/Control bit
IPG	Inter-Packet Gap
LAN	Local Area Network
Link	A full-duplex pair of opposite going channels between the same ends
LV	Low Voltage
LVDS	Low Voltage Differential Signaling
MB	Mega Bytes
Mb	Mega bits
Mbps	Megabit per second
MTU	Maximum Transmission Unit (i.e., maximum packet length)
NRZI	Non-Return-to-Zero, Invert (see the NRZI reference)
O0-O7	Primary port (A-link) output data bits
OB	Primary port (A-link) output Blocking bit
OD	Primary port (A-link) output Data/Control bit
PCB	Printed Circuit Board
Port	A connection of a link to a component
R0-R7	Secondary port (B-link) or LAN input data bits
RB	Secondary port (B-link) or LAN input Blocking bit
RD	Secondary port (B-link) or LAN input Data/Control bit
Rear Panel	Outer rear surface of subrack containing opening(s) for rear-panel modules. The rear panel is often called bulkhead.
Rear-Panel Connectors	Connectors (typically SAN and LAN) mounted on rear-panel module visible to outside of the subrack
Rear-Panel Module	Module mounted in rear-panel opening with interface to the outer world
Res	Reserved signal for future use. This is not a user-defined pin
SAN	System Area Network
S0-S7	Secondary port (B-link) or LAN output data bits (B-link)
SB	Secondary port (B-link) or LAN output Blocking bit
SBC	Single Board Computer
SD	Secondary port (B-link) or LAN output Data/Control bit
Switch	A multiple port device that steers packets among its ports
Symbol	A control character

[This page intentionally left blank.]

M3-SW16-8S

Myrinet-2000 16-Port-Switch Line Card with 8 Serial Ports



These specifications concentrate on the physical characteristics of this product. Please see this Guide to Myrinet-2000 Switches and Switch Networks for information about the installation and the principles of operation of this family of products.

Specifications

Myrinet-2000-Serial ports (8): each 2.0+2.0 Gb/s at a front-panel HSSDC connector to a Myrinet Serial-Link cable up to 10m in length. Each Serial port has a green LED: Off: not connected to an active port; On: connected; Blinking: traffic.

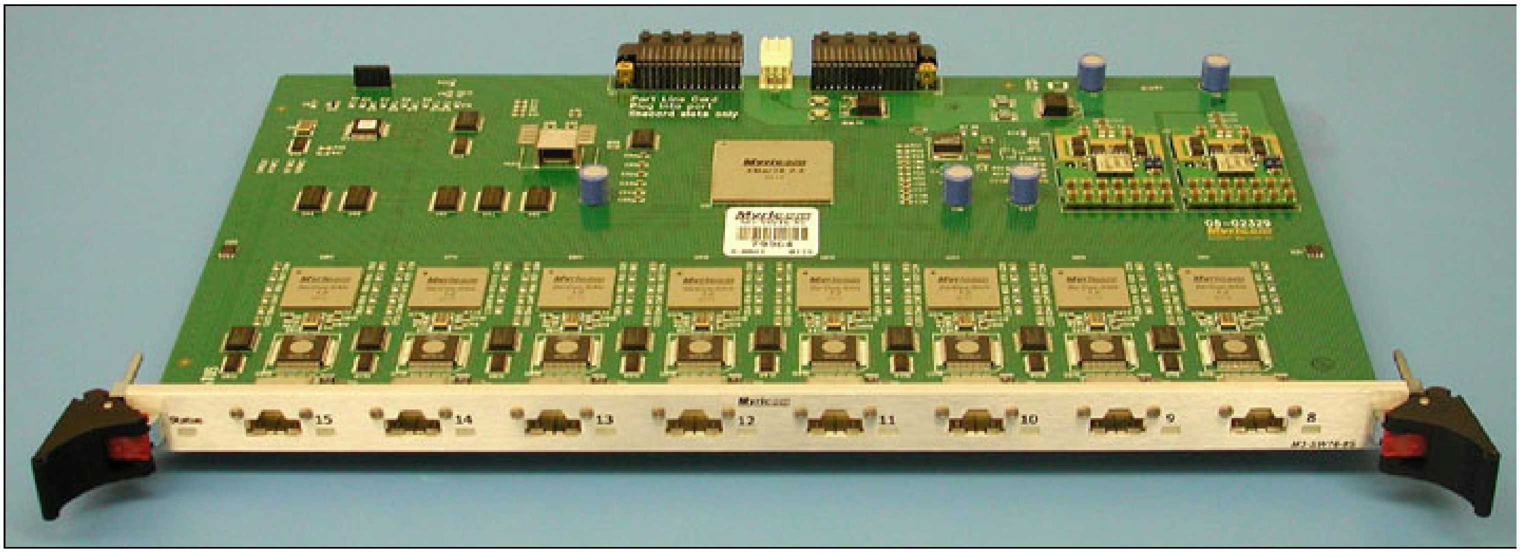
Myrinet-2000-SAN ports (8): each 2.0+2.0 Gb/s at the backplane interface.

Microcontroller functions: Self-test, board identification, voltage monitoring, temperature monitoring (upstream & downstream in the air flow), XBar16 scan path, and status information from the SAN-Serial conversion circuits. All status information is reported on demand through the serial link (to an M3-M monitoring line card). The µC will shut off the line-card power on over-temperature or other fault conditions. The µC controls a Status LED on the front panel: Off: not operating; Green: operating. Yellow: fault.

Power requirements: Dual +12V input power @ 3.1A (37.2W) maximum. This line card is hot-swappable.

Physical characteristics: 9U-220mm card per IEEE 1101.10, with specialized connectorization. Detail dimensional specifications are given here. Weight 20oz (570g).

Regulatory Approvals: When properly installed in a Myricom M3-E16, M3-E32, M3-E64, or M3-E128 enclosure, the line card is fully compliant with EN55024 (1998), EN55022 Class A (1998), VCCI Class A (May 1999), FCC Part 15 Subpart B Class A, CISPR 22/85 Class A, ICES-003 Class A (ANSI C63.4 1992), and AS/NZS 3548 Class A (w/A1 & A2 1997).



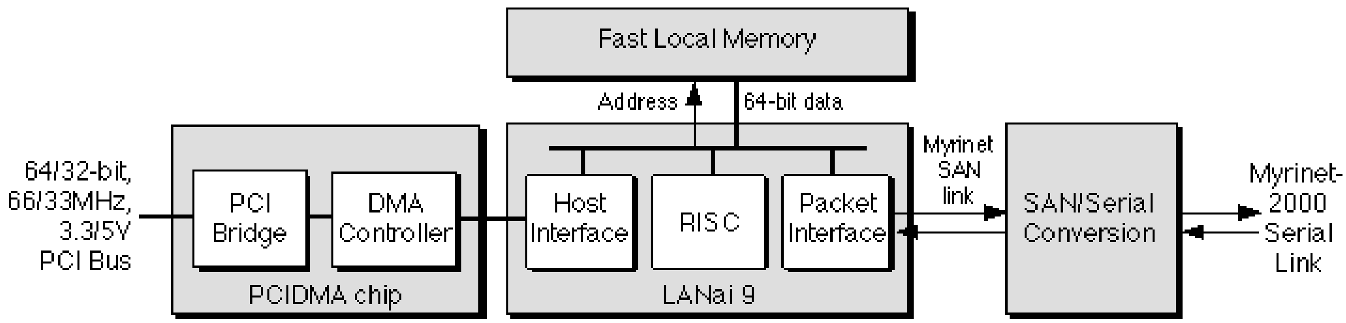
M3-SW16-8S Myrinet-2000 Switch Line Card

Last updated: 18 May 2001

M3S-PCI64B & M3S-PCI64C

Universal, 64/32-bit, 66/33MHz, Myrinet-2000-Serial-Link/PCI interfaces

These universal, 64/32-bit, 66/33MHz, Myrinet-2000-Serial-Link/PCI interfaces are ideal for the most demanding cluster and distributed-computing applications. The interface includes a fast RISC to execute the Myrinet control program, a versatile DMA controller to support zero-copy APIs, and a complete set of high-availability and data-integrity features. The difference between the PCI64B and PCI64C interfaces is the allowed clock rate of the RISC and local memory: 133MHz for the PCI64B, and 200MHz for the PCI64C.



Block diagram

Software Support. All Myricom software support for the PCI64 family of interfaces is based on the GM Myrinet Control Program and the GM API. Software support is available now for:

- Linux on Alpha, Itanium, Pentium, PowerPC, and UltraSPARC
- Win2000 on Itanium and Pentium
- Solaris on Pentium and UltraSPARC
- Tru64 on Alpha
- FreeBSD on Pentium
- NT4 on Pentium
- Irix on O200
- VxWorks on PowerPC

MPICH over GM is also available now.

Specifications

PCI-bus Interface: 64/32-bit, 66/33MHz, supports all burst modes and write-invalidate, master or slave. These interfaces are capable of sustained PCI data rates approaching the limits of the PCI bus (528 MB/s for 64-bit, 66MHz; 264 MB/s for 64-bit, 33MHz or 32-bit, 66MHz; 132 MB/s for 32-bit, 33MHz). However, the data rate to/from system memory will depend upon the host's memory and PCI-bus implementation. These interfaces function correctly in all PCI slots that are compliant with PCI specifications (version 2.2), with either 3.3V or 5V PCI-bus signal levels. (3.3V signaling is required of 66MHz PCI slots, but 33MHz PCI slots may use either 5V or 3.3V signaling.) PCI parity generation and detection is provided. The interface provides a 64-bit Base Address Register (BAR), but will also function properly when programmed with a 32-bit address, per the PCI specifications.

DMA controller: Traverses multiple lists in the interface's local memory to initiate DMA transfers, thus allowing multiple pending DMA operations. In order to support zero-copy APIs efficiently, the DMA operations can be performed with arbitrary byte counts and byte alignments. The DMA controller computes the IP checksum for each transfer. The DMA controller also provides a "doorbell" signalling mechanism that allows the host to write anywhere within the doorbell region, and have the address and data stored in a FIFO queue in the local memory.

Interface processor: LANai 9 RISC operating at up to 133MHz for the PCI64B interfaces, or at up to 200MHz for the PCI64C interfaces. Note: the RISC in the LANai 9 is similar to but is not binary-compatible with earlier LANai RISCs.

Local memory: 2MB (256Kx8B) in the -2 version; 4MB (512Kx8B) in the -4 version. The local memory operates from the same clock as the RISC, i.e., at up to 133 MHz for the PCI64B interfaces, or at up to 200MHz for the PCI64C interfaces. Up to 1,067 MB/s (PCI64B) or 1,600 MB/s (PCI64C) of memory bandwidth is available to support the Myrinet port, the host DMA, and the RISC processor. Byte parity is generated and checked.

Myrinet-2000-Serial port: 2.0+2.0 Gb/s at an HSSDC connector to a Myrinet Serial-Link cable up to 10m in length.

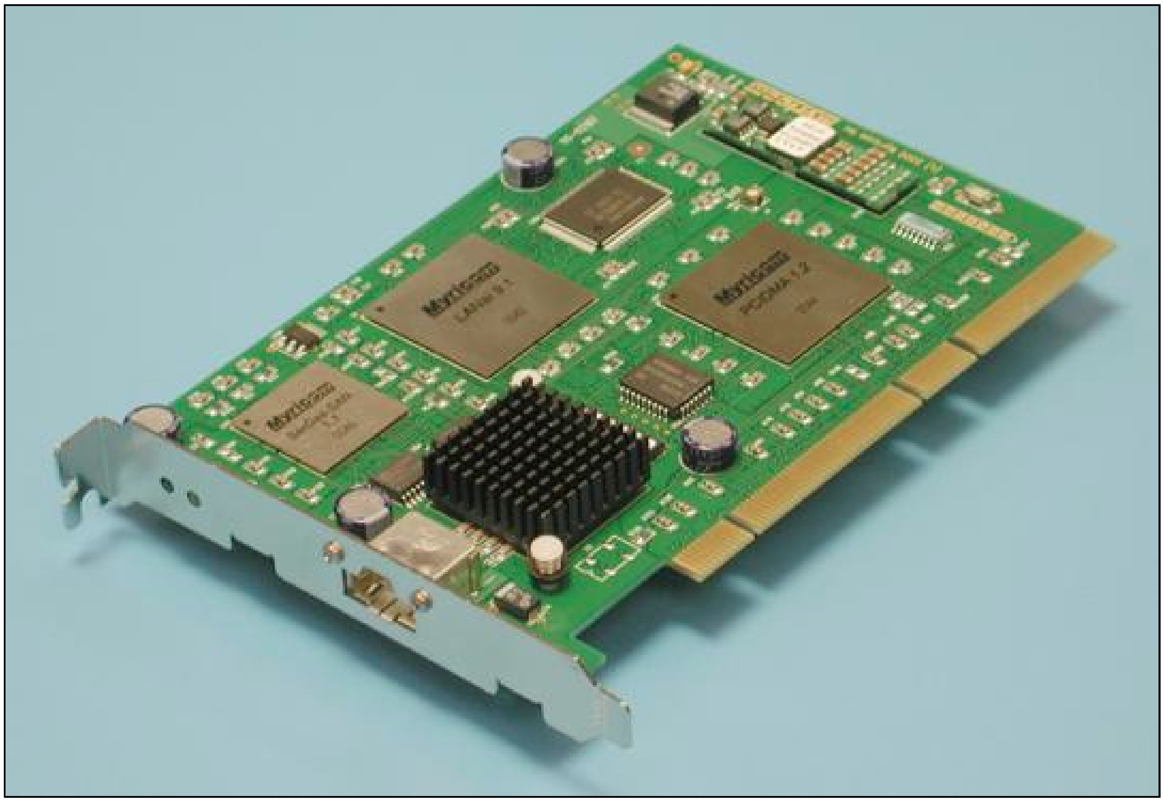
Physical dimensions: PCI Short Card: height 10.7cm, length 18.0cm, total thickness 2.5cm, weight 120g.

Power: The interface is powered from the 5V PCI power: 1.7A (8.5W) maximum for the M3S-PCI64B-2; 2.0A (10.0W) maximum for the M3S-PCI64C-2.

Regulatory Approvals: Fully compliant with EN55024 (1998), EN55022 Class A (1995), VCCI Class A (May 1999), FCC Part 15 Subpart B Class A, CISPR 22/85 Class A, ICES-003 Class A (ANSI C63.4 1992), and AS/NZS 3548 Class A (w/A1 & A2 1997).

Myricom-supported software: Open source, distributed from the Myrinet Software & Customer Support page. These interfaces require the use of the GM software; the MyriAPI software is not available for the PCI64 family of interfaces.

Programmer's Documentation for customers who write their own Myrinet control programs.



M3S-PCI64B-2 Myrinet-2000-Serial/PCI interface

Myricom

Last updated: 13 July 2001